Beschreibung von Proteinbindetaschen für Funktionsstudien und *de Novo-Design* und die Entwicklung von Methoden zur funktionellen Klassifizierung von Proteinfamilien

Dissertation

zur

Erlangung des Doktorgrades

der Naturwissenschaften

(Dr. rer. nat.)

dem

Fachbereich Pharmazie

der Philipps-Universität Marburg

vorgelegt von

Daniel Kuhn

aus Gießen

Marburg/Lahn 2004

Vom Fachbereich Pharmazie der Philipps-Universität Marburg
als Dissertation angenommen am: 20. Dezember 2004

Erstgutachter: Prof. Dr. G. Klebe
Zweitgutachter: Prof. Dr. E. Hüllermeier

Tag der mündlichen Prüfung: 20. Dezember 2004

Die Untersuchungen zur vorliegenden Arbeit wurden auf Anregung von Herrn Prof. Dr. G. Klebe am Institut für Pharmazeutische Chemie des Fachbereichs Pharmazie der Philipps-Universität Marburg in der Zeit von Dezember 2000 bis September 2004 durchgeführt.

Inhaltsverzeichnis

1 Einleitung und Problemstellung

Proteine besitzen auf ihrer Oberfläche Bindetaschen, in denen endogene Liganden oder Substrate erkannt, gebunden und chemisch verändert werden können. Enzyme katalysieren spezifisch einzelne Schritte einer chemischer Reaktionen, was eine fest vorgegebene Anordnung der Reaktionspartner zueinander erfordert. Proteinen aus verschiedenen Organismen, die dieselbe chemische Reaktion katalysieren, weisen oft ein sehr ähnliches Muster an molekularen Erkennungseigenschaften auf, das die räumliche Ausrichtung der Reaktionspartner aufeinander festlegt. Daher liegt der Schluß nahe, daß ein so definiertes Muster von Erkennungseigenschaften in Bindetaschen mit einer bestimmten Funktion eines Proteins einhergeht. Proteine vergleichbarer Funktion sollten daher, unabhängig von dem weiteren Faltungs- und Sequenzmuster des betrachteten Proteins, auch ein verwandtes Muster molekularer Erkennungseigenschaften in ihren Bindetaschen abbilden.

Das Wissen über Proteinsequenzen und -strukturen wächst derzeitig stark an. Nach der initialen Entschlüsselung des humanen Genoms [Lander et al., 2001; Venter et al., 2001] und nach dem Abschluß dieser Arbeiten [Consortium, 2004] im Sommer 2004 sind nun ca. 20000 bis 25000 Gensequenzen bekannt. Neben diesem immensen Informationsgehalt haben methodische Fortschritte auf dem Gebiet der Proteinstrukturaufklärung (High-Throughput Kristallographie [Blundell et al., 2002], NMR [Pellecchia et al., 2002]) dafür gesorgt, daß das strukturelle Wissen über Proteinstrukturen ebenfalls stark zugenommen hat. Die Zahl der öffentlich zugänglichen 3D-Strukturen von Proteinen in der Proteindatenbank PDB [Berman et al., 2000] ist in den letzten Jahren exponentiell angestiegen. Die PDB umfasst zur Zeit über 27.000 Strukturen (Stand September 2004). Im Zuge von *Structural Genomics* Projekten wird versucht, experimentell alle möglichen Proteinfaltungsmuster strukturell aufzuklären. Daraus resultieren Proteinstrukturen, deren Funktion zur Zeit der Strukturaufklärung noch nicht bekannt ist. Die Fähigkeit, die Funktion von diesen Proteinen aus ihrer Struktur abzuleiten, kann für die Annotierung sehr nützlich sein und ergänzt bestehende Ansätze, die Sequenz- und Faltungshomologien zur Funktionsbestimmung benutzen.

Die moderne pharmazeutische Forschung beginnt mit der Identifizierung eines molekularen Targets (Rezeptor, Protein, Enzym, Ionenkanal), das ursächlich mit dem Entstehen oder der Progression einer Krankheit verbunden ist. Heutzutage bekann-

te Arzneistoffe greifen an ungefähr 500 verschiedenen Zielmolekülen an [Drews, 2000]. Man nimmt an, daß die Zahl der Targets im Genom noch wesentlich höher liegt. Die Herausforderung besteht nun darin, aus den bekannten Genen, die Genprodukte als Targets zu identifizieren und zu validieren [Wang et al., 2004], deren Funktion mit kleinen Molekülen beeinflußt und moduliert werden kann - sogenannte *druggable targets* [Hopkins and Groom, 2002]. Eine vergleichende Analyse von Proteinbindetaschen kann im Identifizieren von Kriterien, die entscheidend für die Adressierbarkeit von Bindetaschen mit kleinen Molekülen sind, helfen.

Ist ein Zielprotein als ein valides Target identifiziert, werden in der Regel niedermolekulare Liganden entwickelt, die, um eine möglichst hohe Spezifität und Selektivität zu erreichen, eine spezifische Bindung zum Protein ausbilden sollen. Diese Bindung geschieht meistens in einer definierten Bindetasche. Existieren andere Proteine mit vergleichbaren molekularen Erkennungseigenschaften in ihren Bindetaschen, so kann es auch dort zu einer Bindung kommen. Daraus kann man zwei Dinge ableiten. Zum einen können verschiedene Liganden identifiziert werden, die in verschiedene Proteine mit ähnlichen Bereichen in der Bindetaschen binden. Binden Ligandenfragmente in Subtaschen, die physikochemisch ähnliche Eigenschaften aufweisen, werden dadurch wertvolle Ideen für das Design und den Austausch von bioisosteren Verbindungen geliefert. Zum anderen können sich Ähnlichkeiten in den Bindetaschen von zwei Proteinen in einer unerwünschten Nebenwirkung manifestieren. Diese Kreuzreaktivität in anderen Proteinen ist dann einfach vorherzusagen, wenn die betrachteten Proteine zu einer gemeinsamen Proteinfamilie gehören, d.h. untereinander hohe Sequenz- und Faltungshomologie besitzen und auch eine große Ähnlichkeit im Aufbau der Bindetasche festzustellen ist. Solche Kreuzreaktivitäten sind deutlich schwieriger abzuschätzen, wenn beide Proteine keine Sequenz- oder Faltungshomologie aufweisen. Derzeit praktisch noch unmöglich ist eine Vorhersage für den Fall, daß diese Proteine, neben den ähnlichen Bindungsepitopen in ihren Bindetaschen, keinerlei weitere, einfach zu erkennende Homologie untereinander besitzen.

In der modernen Wirkstoffentwicklung gewinnen zunehmend Ansätze an Bedeutung, die nicht nur Eigenschaften von einzelnen Proteinen, sondern von kompletten Proteinfamilien untersuchen [Naumann and Matter, 2002]. Diese Ansätze werden auch als *Chemogenomics* [Jacoby et al., 2003] bezeichnet. Dahinter steckt die Grundannahme, daß ähnliche Proteine ähnliche Liganden binden sollten. Durch die Analyse von Ähnlichkeiten und Unterschieden in den Bindestellen einer Proteinfamilie lassen sich Bereiche, die für Spezifität und Selektivität von Liganden entscheidend sind, identifizie-

ren. Aufbauend auf den Ähnlichkeiten in den Bindestellen von Proteinen einer Familie läßt sich eine Klassifizierung von diesen Proteinfamilien erstellen, die sich auf funktionell wichtige Bereiche im Protein konzentriert. Diese Klassifizierung ergänzt andere Klassifizierungen, die beispielsweise auf Struktur-Affinitätsdaten von kleinen Molekülen beruhen [Frye, 1999].

In einer vorangegangen Arbeit [Schmitt, 2000] wurde eine Datenbank entwickelt, die Informationen über Proteinbindetaschen effizient verwaltet und Ähnlichkeitssuchen mit ausgewählten Proteinbindetaschen erlaubt. Ziel dieser vorliegenden Arbeit ist es, diese Datenbank auszubauen und die Bindetaschenvergleiche zu optimieren, so daß Ähnlichkeitssuchen in großen Datensätzen möglich werden. Dadurch sollen folgende Themenschwerpunkte, die inhaltlich alle mit dem Aufbau und den physikochemischen Eigenschaften von Proteinbindetaschen und den Ähnlichkeitsbeziehungen zwischen ihnen verknüpft sind, bearbeitet werden:

- **Funktionsanalyse von Proteinstrukturen**, deren Funktion noch nicht bekannt ist. Gerade Proteinstrukturen, die in *Structural Genomics*-Initiativen aufgeklärt werden, stellen interessante Testfälle für die Funktionsannotation dar.
- **Charakterisierung von Proteinbinde-(sub)taschen nach den Liganden oder Ligandfragmenten**, die in diese Taschen binden. Dabei soll der Frage nachgegangen werden, ob man das Wissen über Ähnlichkeiten in der Bindetasche im Design von neuen Liganden nutzen kann. Bindetaschen, die ein ähnliches physikochemisches Umfeld aufweisen, sollten in der Lage sein, ähnliche Liganden zu binden.
- **Suche nach Kreuzreaktivitäten als Ursache für Nebenwirkungen**. Lassen Ähnlichkeiten im strukturellen Aufbau und in den Eigenschaften der Bindetaschen eine Verbindung zwischen nicht verwandten Proteinen herstellen, die auf eventuell experimentell beobachtbare Kreuzreaktivitäten schließen lassen?
- **Klassifizierung von Proteinfamilien anhand der Eigenschaften ihrer Bindetaschen**. Es existieren eine Reihe von Verfahren, die Proteine oder Proteinfamilien anhand ihrer Sequenz- oder Faltungsmusterähnlichkeiten klassifizieren. Ein entsprechendes Konzept, das eine solche Klassifikation für Bindetaschen aufstellt ist noch nicht bekannt. Interessant wäre es, die Ergebnisse einer Klassifikation basierend auf Ähnlichkeiten und Unähnlichkeiten in der Bindetasche existierenden Klassifikationen gegenüberzustellen.

In Kapitel 2 werden Ansätze aus der Literatur, die Proteine anhand Sequenz- und/oder Faltungsinformation vergleichen, vorgestellt. Kapitel 3 geht auf Erweiterungen und Optimierungen in Cavbase ein. Kapitel 4 enthält die Ergebnisse der Ähnlichkeitssuchen und Funktionsannotierungsstudien mit Cavbase. Die Ergebnisse der funktionellen Klassifizierung von Proteinfamilien am Beispiel der Proteinfamilien der Carboanhydrasen und Proteinkinasen werden in Kapitel 5 dargestellt und diskutiert. Kapitel 6 stellt die Ergebnisse eines Subtaschen-gesteuerten Optimieren und Charakterisieren einer Protease vor. Kapitel 7 fasst die Ergebnisse der Arbeit zusammen und gibt einen Ausblick.

2 Ansätze zum Vergleich von Proteinstrukturen aus der Literatur

In dem folgenden Kapitel sollen Ansätze aus der Literatur zum Vergleich und zur Struktur- und Funktionsvorhersage vorgestellt werden (zur Übersicht siehe auch Tabelle 2.1). Dieser ausführliche Vergleich soll zeigen, wo sich unser Ansatz von anderen Methoden abhebt und weshalb damit Lösungsansätze zu erwarten sind, die mit keinen in der Literatur beschriebenen Methoden derzeit geleistet werden können. Die Analyse von Verwandtschaftsbeziehungen zwischen Proteinen kann auf verschiedenen Ebenen durchgeführt werden. So kann bereits auf primärer Ebene durch den Vergleich der Abfolge der Aminosäurebausteine (Sequenz) mehrerer Peptidketten untereinander auf eine eventuell vorhandene Verwandtschaft zwischen Proteinen geschlossen werden. Ebenso können Faltungsmustervergleiche hilfreiche Informationen zur Klassifizierung unbekannter Proteinstrukturen liefern. Am anspruchvollsten sind Algorithmen, die nach gemeinsamen dreidimensionalen Substrukturen suchen und so unter Umständen zu noch genaueren Informationen über die Funktion eines Proteins gelangen. Bestehende Ansätze lassen sich grob nach den verwendeten Informationen in drei Kategorien gliedern:

- Sequenzvergleiche
- Vergleich von Faltungsmustern und Sekundärstrukturelementen
- Vergleich von 3D-Substrukturen

2.1 Sequenzvergleiche

Zeigen zwei Proteine eine hohe Sequenzidentität, dann kann von ihr aus auf eine strukturelle und eventuell auch funktionelle Ähnlichkeit geschlossen werden [Chothia and Lesk, 1986; Wood and Pearson, 1999; Wilson et al., 2000; Gan et al., 2002]. Sequenzvergleiche zur Funktionszuweisung von Proteinen werden heute standardmäßig durchgeführt und es existieren eine Reihe von Algorithmen, die nach Ähnlichkeiten in Sequenzen suchen. Einige Algorithmen versuchen Sequenzen entlang ihre kompletten Länge

optimal zu überlagern (globales Alignment), modifizierte Varianten sind mehr auf die Suche nach lokalen Alignments ausgelegt. Die bekanntesten Verfahren für beide Algorithmen sind von Needleman und Wunsch [Needleman and Wunsch, 1970] und von Smith und Waterman [Smith and Waterman, 1981] vorgestellt worden. Beide Verfahren benutzen dynamische Programmierung, um ein optimales Alignment von zwei Sequenzen zu erhalten. Während der Algorithmus von Needleman und Wunsch ein optimales globales Alignment zweier Sequenzen liefert, findet die von Smith und Waterman abgewandelte Methode optimale lokale Überlagerungen. Beide Ansätze sind aber relativ rechen- und speicherintensiv, so daß vielfach heuristische Methode wie FASTA [Pearson and Lipman, 1988; Pearson, 1990] und BLAST [Altschul et al., 1990] zum Einsatz kommen. Diese Verfahren finden nicht in allen Fällen das optimale, globale Alignment, liefern aber im allgemeinen gute Ergebnisse. Weiterentwicklungen von BLAST, wie gapped-BLAST (Berücksichtigung von Lücken) und PSI-BLAST [Altschul et al., 1997] (Benutzung einer Positions-sensitiven Bewertungsmatrix) haben die Sensitivität und Selektivität wesentlich erhöht, so daß auch entfernt verwandte Proteine noch als ähnlich detektiert werden können. Die Fähigkeit von BLAST einen quantitativen Signifikanzwert zu jedem Sequenzvergleich zu liefern, erleichtert wesentlich die Beurteilung der gefundene Ergebnisse. Grosse Sequenzdatenbanken (SWISS-PROT [Boeckmann et al., 2003], Genbank [Benson et al., 2003]) können standardmäßig mit diesen Verfahren nach ähnlichen Sequenzen durchsucht werden.

Sinkt die Sequenzidentität unter 20-25 % ist man mit einfachen paarweisen Sequenzvergleichen nicht mehr in der Lage entfernte Verwandtschaftsbeziehungen festzustellen [Park et al., 1998]. Sensitivere Methoden wie Profil Ansätze [Gribskov et al., 1987, 1990] oder Hidden Markov Modelle (HMM) [Krogh et al., 1994; Eddy, 1996] sind nötig, um diese Ähnlichkeiten noch zu detektieren und werden so zur Klassifizierung von Proteinen und zur Funktionszuweisung genutzt. Profil-basierte Analysen nutzen Informationen aus einem multiplen Alignment, um Muster in dem Sequenzalignment homologer Strukturen zu entdecken. Zum Auffinden eines optimalen Alignments werden sogenannte Positions-spezifische Bewertungsmatrizen (PSSM) verwendet, die konservierten Positionen höher bewerten als abweichende Positionen. HMM sind statistische Modelle des multiplen Sequenzalignments. Sie beschreiben die Positions-spezifische Information, die in den einzelnen Spalten eines multiplem Sequenzalignments verborgen ist und berücksichtigen auch Lücken im Alignment. HMM haben den Vorteil, daß sie ausschließlich Sequenzinformationen benutzen, ein multiples Sequenzalignment muss

beim Aufstellen eines HMM nicht vorliegen. HMM sind daher auch allgemeiner anwendbar als Profile.

Eine ganze Reihe von Motivdatenbanken, wie PROSITE [Bucher and Bairoch, 1994; Falquet et al., 2002], Pfam [Sonnhammer et al., 1998; Bateman et al., 2002], BLOCKS [Henikoff and Henikoff, 1996; Henikoff et al., 1999, 2000], PRINTS [Attwood, 2002; Attwood et al., 2003], etc. benutzen Profile oder Hidden Markov Modelle, um nach Sequenzmotiven zu suchen und sind so in der Lage, entfernte Verwandtschaften festzustellen.

2.2 Vergleich von Faltungsmustern und Sekundärstrukturelementen

Obwohl man mit der Anwendung von Sequenzvergleichsverfahren auch eine ganze Reihe von entfernten Ähnlichkeiten (Sequenzidentität < 35 %) zwischen Proteinen detektieren kann, ist man in einigen Fällen auf den Einsatz von dreidimensionalen Vergleichsverfahren angewiesen. Methoden zur Klassifizierung und Identifizierung von strukturellen Verwandtschaften zwischen Proteinen auf der Basis von globalen Faltungsähnlichkeiten sind weit verbreitet. Aus ihnen sind hierarchische Systeme entstanden, die Proteine nach ihrer Faltung, ihren entwicklungsgeschichtlichen Zusammenhängen oder nach ihrer Funktion einteilen. Diese Systeme arbeiten entweder automatisch oder sind auf manuelle Intervention angewiesen. Vielen der Klassifikationsschemata ist gemein, daß sie Proteine aus Domänen (räumlich getrennte Strukturen, die sich alleine falten und ihre Funktion ausüben) aufgebaut verstehen [Ponting and Russell, 2002] und sie nach den Eigenschaften der Domänen einteilen. Wichtige Klassifizierungsschemata sind: SCOP [Murzin et al., 1995; Lo Conte et al., 2002], CATH [Orengo et al., 1997; Pearl et al., 2000], FSSP/DALI [Holm and Sander, 1994, 1996a,b, 1998; Dietmann et al., 2001], MMDB [Gibrat et al., 1996]. Davon setzen sich die ENZYME-Datenbank [Bairoch, 2000] und die Datenbank BRENDA [Schomburg et al., 2002b,a] ab, die Proteine anhand der katalysierten Reaktion einteilen.

Die SCOP (Structural Classfication of Proteins) klassifiziert Proteine nach ihren strukturellen und entwicklungsgeschichtlichen Verwandtschaften. Diese Klassifizierung wird durch visuelle Inspektion und durch Vergleich der Proteinstrukturen mit einer Reihe von automatischen Methoden erreicht. Der SCOP-Ansatz definiert fünf hierarchische

Ebenen. Proteindomänen werden zu Familien zusammengefasst, wenn sie eine ausreichende Sequenzidentität zeigen oder eine ähnliche biochemische Reaktion katalysieren. Superfamilien haben einen gemeinsamen Ursprung. Superfamilien werden zu Faltungen zusammengefasst, wenn ähnliche Sekundärstrukturmuster in einer spezifischen Topologie angeordnet sind. Die oberste Ebene bilden fünf Faltungsklassen, je nach Anteil von α-Helices und β-Faltblättern.

Die CATH Datenbank ist ein hierarchisches Klassifizierungsschema, das Proteindomänen auf vier Hauptebenen einteilt: Class(C), Architecture(A), Topology(T) and Homologous superfamily (H). Die Klasse wird über den Gehalt an Sekundärstrukturelemente (SSE) automatisch, die Architektur, die grob die räumliche Anordnung der SSEs beschreibt, wird manuell zugewiesen. Die Topologie bewertet, wie SSEs zueinander angeordnet sind und Proteine werden zu homologen Familien aufgrund von Sequenz- und Faltungsähnlichkeiten zusammengefasst.

Die FSSP (**F**old classification based on **S**tructure-**S**tructure alignment of **P**roteins) Datenbank enthält einen vollständigen Vergleich aller Proteinstrukturen in der Protein Data Bank (PDB). Die Vergleiche werden mit dem Programm DALI automatisch durchgeführt. DALI ermittelt ähnliche Strukturen anhand von Cα-Atom-Abstandsmatrizen. Anhand von Sequenz- und Strukturähnlichkeiten werden die Proteinketten in einen repräsentativen und homologen Datensatz eingeteilt.

Die vom NCBI (**N**ational **C**enter for **B**iotechnology **I**nformation) herausgegebene annotierte Strukturdatenbank MMDB (Molecular Modeling DataBase) enthält eine repräsentative Untermenge der PDB. Proteinstrukturen werden mit Hilfe halbautomatischer Methoden in eine Taxonomie eingeordnet. Die MMDB enthält zusätzliche Annotationen zu den einzelnen Proteinfamilien. Strukturelle Vergleiche werden mit dem VAST (Vector Alignment Search Tool) Programm durchgeführt. VAST vergleicht Proteine aufgrund der Ähnlichkeit ihrer Sekundärstrukturelemente.

Die ENZYME Datenbank klassifiziert Enzyme nach den Reaktionen, die sie katalysieren und liefert wertvolle Informationen für eine funktionelle Annotation von Proteinen. Sie folgt den Empfehlungen des Nomenclature Kommittees der International Union of Biochemistry and Molecular Biology (IUBMB). Zur Zeit sind 4173 Enzyme in der Datenbank enthalten (Juni 2003). Die ENZYME Datenbank charakterisiert Enzyme durch vier Nummern. Die erste teilt das Enzym je nach katalysierten Reaktion in sechs Basisklassen ein (Oxidoreduktasen, Transferasen, Hydrolasen, Lyasen, Isomera-

sen und Ligasen), die zweite Nummer hängt von der Bindung oder dem Substrate ab, auf die/das das Enzym wirkt, durch die dritte Nummer werden Substrate/Produkt Spezifitäten beschrieben und die vierte Nummer wird nach Abhängigkeit von Kofaktoren zugewiesen.

Die Datenbank BRENDA (BRaunschweig ENzyme DAtabase) enthält eine Fülle an funktionellen, biochemischen und metabolischen Informationen über Enzyme, die durch die manuelle Auswertung von über 45000 Literaturstellen zusammengestellt worden sind.

Es existieren eine Vielzahl an Ansätzen, Proteine strukturell zu überlagern (siehe Übersichtsarbeiten [Gibrat et al., 1996; Godzik, 1996; Eidhammer et al., 2000; Koehl, 2001; Orengo et al., 2001; Carugo and Pongor, 2002]). Modelliert werden Proteinstrukturen durch die Position ihrer Cα-Atome, durch Fragmente von Aminosäuren oder durch Sekundärstrukturelemente. Oft werden diesen Deskriptoren noch Eigenschaften wie dreidimensionale Koordinaten, Winkelabhängigkeiten, Aminosäuretyp, Sekundärstrukturtyp, Kurvigkeit oder Krümmung der Proteinoberfläche zugeordnet und im Vergleichsprozess verwendet. Aus algorithmischer Sicht werden Graph-basierte Verfahren, Geometric Hashing, doppelte dynamische Programmierung, Vergleich von Distanzmatrizen und Fragment-basierte Ansätze verwendet.

2.3 Vergleich von 3D-Substrukturen

Funktionelle Untersuchungen haben gezeigt, daß nur eine geringe Korrelation zwischen Faltung und Funktion von Proteinen besteht [Martin et al., 1998; Todd et al., 2001; Orengo et al., 2001; Nagano et al., 2002; Anantharaman et al., 2003]. Es gibt Faltungsmuster, die eine Vielzahl an unterschiedlichen Reaktionen katalysieren (z.B. TIM-Barrel, β-Propeller-Faltung). Proteine können aber auch selbst dann noch eine ähnliche Funktion besitzen, wenn sie keinerlei Sequenz- und Faltungshomologie zeigen (Serinproteasen aus der Trypsin bzw. Subtilisinfamilie oder Chorismatmutase zweier Spezies aus *E. coli* und *S. cerevisiae*). Methoden, die Proteine nur aufgrund ihres Faltungsmusters vergleichen, können solche Ähnlichkeiten nicht detektieren. Verfahren, die in Proteinen nach ähnlichen Substrukturen suchen, sind in der Lage, in diesen Fällen Ähnlichkeiten zu entdecken. Im Folgenden sollen einige Ansätze, die nach gemeinsamen Substrukturen in Proteinen suchen, vorgestellt werden. Diese Verfahren werden noch einmal

danach unterschieden, ob sie zur Suche sogenannte Template (räumlich definierte Muster aus drei oder mehreren Aminosäuren) benutzen oder ob sie ohne solche Vorgaben auskommen.

2.3.1 Vergleiche unter Verwendung von Templaten

Katalytisch wichtige Aminosäuren sind nicht selten auf Sequenzebene wie auch strukturell in Proteinfamilien in allen Vertretern dieser Familie zu finden. Durch die Suche nach 3D-Äquivalenten zu solchen Sequenzmotiven sollte man in der Lage sein, funktionelle Verwandtschaften zu entdecken. Verschiedene Ansätze greifen diese Idee auf, indem sie in Proteindatenbanken nach dem Vorhandensein von zuvor definierten oder automatisch generierten Templaten suchen. Man hofft so, die für die Funktion entscheidende Reste zu identifizieren (z. B. die katalytische Triade in Serinproteasen). Die Beschränkung auf relativ kleine Suchtemplate erlaubt außerdem das schnelle Durchmustern großer Datenmengen.

Artymiuk et al. beschreiben eine Methode (ASSAM), um nach 3D-Seitenkettenmotiven in Datenbanken zu suchen. Ein algorithmisch ähnlicher Ansatz wurde bereits von ihnen benutzt, um Proteine auf der Ebene von Sekundärstrukturelementen zu vergleichen [Grindley et al., 1993]. Die Seitenketten der Aminosäuren werden durch zwei Pseudoatome, die den funktionellen Teil der Aminosäure beschreiben, repräsentiert. Ein Seitenkettenmotiv setzt sich aus mehreren Aminosäuren und einer Distanzmatrix zwischen allen Pseudoatomen zusammen und umfasst typischerweise drei Aminosäuren. Zusätzlich können zur Definition eines Seitenkettenmotivs einige generische Aminosäurentypen definiert werden (sauer, basisch, Glu/Gln, etc.). Das Protein wird als ein gewichteter Graph mit den Pseudoatomen als Knoten (ausgestattet mit dem Aminosäuretyp) und den Distanzen zwischen den einzelnen Pseudoatomen als Kanten dargestellt. Ein Ullmann-Subgraph Algorithmus wird verwendet, um alle Subgraphen zu permutieren und um so die An- oder Abwesenheit eines bestimmten Musters in einem Graphen zu erkennen. An vier Beispielen ist die Methode validiert worden und es konnten ähnliche 3D-Motive zuverlässig wiedererkannt werden. Der Ansatz bringt aber im Vergleichsprozess nur identische Aminosäuren zur Deckung, außerdem werden Hauptkettenatome vernachlässigt.

Spriggs et al. [Spriggs et al., 2003] haben einige Verbesserungen am ursprünglichen Verfahren vorgenommen. So verfügt ASSAM über eine größere Zahl von generischen

Aminosäuretypen, die allgemeinere Suchen möglich machen. Eine Reihe von optionalen Heuristiken erhöhen die Selektivität der Ähnlichkeitsanfragen, indem sie den Suchraum einschränken. Dabei werden u. a. nur solche Aminosäuren als ähnlich angesehen, die eine vergleichbare Distanz zu einer Bindestelle (definiert durch die Anwesenheit von einem Heteroatom) besitzen, aus einem vergleichbaren lokale Faltungsmotiv entstammen, oder eine ähnliche Solvenszugänglichkeit aufweisen. Als wichtige Neuerung ist die Berücksichtigung der Interaktionsmöglichkeiten der Peptidhauptkette zu nennen. Als Testbeispiel diente die katalytische Triade der Serinproteasen am Beispiel des Chymotrypsins.

Das von Wallace et al. entwickelte Programm TESS (TEmplate Search and Superposition) ist ein alternatives Verfahren zur Suche nach vorgegebenen 3D-Mustern. Es verwendet einen Geometric Hashing Algorithmus, um ähnliche Anordnungen von Aminosäuren zu ermitteln. Ein repräsentativer Datensatz von 639 Strukturen wurde mit drei definierten Motiven aus verschiedenen Proteinfamilien (u.a. katalytische Triade der Serinproteasen) nach Ähnlichkeiten durchsucht. Anhand der RMSD-Werte der gefundenen Lösungen konnte zwischen katalytischen und nicht-katalytischen Triaden unterschieden werden und Serinproteasen danach klassifiziert werden [Wallace et al., 1996, 1997]. Aus diesen Arbeiten ist die Datenbank PROCAT entstanden, die manuell zusammengestellte Informationen über katalytische Reste in aktiven Zentren enthält und zur Änhlichkeitssuche verwendet werden kann [Wallace et al., 1996, 1997; Bartlett et al., 2002].

Ein interessantes Verfahren zur Detektion von Seitenkettenmotiven wird von Russell [Russell, 1998a] vorgestellt. Die Aminosäuren werden hier durch ihre Cα-, Cβ-Atome und ein an der Seitenkette lokalisiertes 'funktionelles Atom' dargestellt. Durch multiple Sequenzvergleiche von Proteinen gleicher Strukturklasse werden funktional wichtige Aminosäuren identifiziert. Untersuchungen über Aminosäureverteilungen in aktiven Zentren [Bartlett et al., 2002] zeigen, daß diese sehr ähnlich angeordnet und daß aliphatische Aminosäuren als katalytische Aminosäuren unterrepräsentiert sind. Aus diesem Grund werden aliphatische Reste (AFGILPV) und in einer Familie nicht konservierte Aminosäuren beim Vergleich nicht berücksichtigt. Ein 'depth-first' Suchalgorithmus detektiert alle räumlich ähnlichen Anordnungen identischer Motive. Aufgefundene Muster werden anhand eines gewichteten RMSD-Werts beurteilt. Interessant ist außerdem, daß Russell ausgehend vom RMSD Wert in der Lage war, eine statistische Signifikanz für die einzelnen Überlagerungen zu berechnen, um so die erhaltenen Lösungen besser beurteilen zu können. Das Verfahren wurde an einer Vielzahl von Beispielen getestet.

Dabei konnten auch einige potentielle Ähnlichkeiten nicht strukturhomologer Rezeptoren entdeckt werden.

Kürzlich haben Stark und Russell den Datenbankserver Pints (Patterns in Nonhomologous Tertiary Structures) [Stark et al., 2003; Stark and Russell, 2003] vorgestellt, der das Verfahren von Russell [Russell, 1998a] implementiert und verschiedene Arten von Suchanfragen zulässt. So sind Vergleiche von Proteinen gegen eine Datenbank von Aminosäuremotiven, der Vergleich eines Motivs gegen eine Datenbank von Proteinstrukturen und komplette Proteinstrukturvergleiche möglich. Motive können aus bis zu elf Aminosäuren bestehen. Die statistische Signifikanz von RMSD-Abweichungen einer lokalen Ähnlichkeit in zwei Proteinstrukturen kann abgeschätzt werden. Aus zufälligen RMSD-Verteilungen verschiedener Aminosäuremotive lässt sich ein Erwartungswert ableiten, mit dem man in der Lage ist - unabhängig von der Größe der gesuchten Substruktur - signifikante Treffer vom Hintergrundrauschen abzusetzen.

Einen sehr ähnlichen Ansatz wie Russell [Russell, 1998a] verwendet Hamelryck [Hamelryck, 2003] zur Identifizierung von Seitenkettenmotiven. Triaden aus Aminosäuren werden als Suchmotive verwendet und in einer multidimensionalen Indexstruktur abgelegt und gesucht. Ähnlich der Aminosäurenrepräsentation von Russel werden rein hydrophobe Aminosäuren nicht berücksichtigt. Die Aminosäuren werden durch eine Untermenge an funktionell wichtigen Atomen kodiert. Triaden aus Aminosäuren werden räumlich anhand der Abstände zwischen diesen Atomen beschrieben. Diese Charakterisierung erlaubt auch das Auffinden von spiegelbildlichen Geometrien. Die Signifikanz eines gefundenen RMSD-Wertes einer Überlagerung zweier Triaden kann durch eine experimentell gewonnene Verteilungsfunktion abgeschätzt werden. Neben klassischen Validierungsbeispielen wurde eine Liste von SCOP repräsentativen Proteinen auf interessante Triaden untersucht.

Wangikar et. al. [Wangikar et al., 2003] stellen ein Verfahren (DRESPAT) vor, um nach Seitenkettenmotiven in Proteinfamilien zu suchen. Die Repräsentation der Aminosäuren erfolgt auf dieselbe Weise wie bei Russell [Russell, 1998a]. Detektiert werden strukturelle Muster, die aus drei bis sechs Aminosäuren bestehen. Ein Cliquealgorithmus enumeriert alle möglichen ähnlichen strukturellen Motive (zwischen 2000 und 50000 pro Protein) in einem Protein. Anschließend wird nach gleichen Mustern in zwei Proteinen gesucht, zusätzlich können in einem zweiten Lauf des Cliquealgorithmus ähnliche Anordnungen von Mustern in Proteinfamilien detektiert werden. Einfache Statistikauswertungen versuchen einen Signifikanzwert über die Anzahl der gefundenen Muster einer

bestimmten Größe (Dreieck, Tetraeder, Hexaeder) in mehreren Proteinen abzuschätzen. Der Erwartungswert ist aus der statistischen Analyse von SCOP-Superfamilien (diese Proteine weisen in der Regel eine Ähnlichkeit zueinander auf) abgeleitet und berücksichtigt wie oft ein Muster einer bestimmten Größe in einer Menge von untersuchten Proteinen gefunden wird. Als Validierungsdatensatz dienten 17 Familien aus 128 repräsentative SCOP Superfamilien mit mindestens 10 Einträgen. Innerhalb einer Superfamilie entsprach das am höchsten bewertete Motiv in den meisten Fällen dem katalytischen.

Kleywegt [Kleywegt, 1999] folgt einem Ansatz, der nach einem bestimmten Motiv in Proteinstrukturen sucht (SPASM) oder ein einzelnes Protein gegen eine Datenbank von Motiven (RIGOR) durchmustern kann. Die Aminosäuren werden durch ihr Cα-Atom und ein am Schwerpunkt der Seitenkettenatome lokalisiertes Pseudoatom dargestellt. Als ähnlich angesehen werden hier gleiche Aminosäuren, es besteht aber die Möglichkeit verschiedene Aminosäuren nur durch ihre Cα-Atome zur Übereinstimmung zu bringen. Für ein gesuchtes Motiv werden zuerst alle möglichen Treffer in einem Protein aus der Datenbank mit Proteinstrukturen generiert, um dann über Distanzvergleiche gute Überlagerungen zu finden. Dieser Algorithmus wurde bereits in das Faltungsmustererkennungsprogramm DEJAVU [Kleywegt et al., 1994] implementiert. An vier Beispielen zum Auffinden von Proteinen ähnlicher Funktion (Retinol bindendes Protein, Cellobiohydrolase I und Thermolysin, Auffinden von linksgängigen Helices) ist das Verfahren validiert worden.

Ähnlichkeitssuchen mit Templaten sind fähig, ähnliche Motive in Proteinen zuverlässig auch in nicht verwandten Proteinen zu detektieren. Einige der genannten Methoden sind über Signifikanzabschätzungen außerdem in der Lage, zwischen katalytischen und nicht-katalytischen Triaden zu unterscheiden. Nachteile dieser Methoden liegen in der schon vorgegebenen strukturellen Formulierung eines Suchmotivs, die die zu erwartende Ähnlichkeit schon vorwegnehmen. Die Beschränkung auf relativ kleine Template erlaubt nicht das Detektieren ganzer Bindetaschenbereiche. Andererseits können dadurch aber Signifikanzabschätzungen getroffen werden, die wichtige Motive von zufälligen unterscheiden können. In den vorgestellten Ansätzen werden die Aminosäuren oft nur durch wenige Deskriptoren beschrieben.

2.3.2 Vergleiche ohne Verwendung von Templaten

Als zweite größere Gruppe zum Substrukturvergleich von Proteinen sind die Verfahren zu nennen, die ohne Definition von vorgegebenen Templaten auskommen. Zur Ähnlichkeitssuche werden ganze Proteine oder Substrukturen verwendet.

Aus der Gruppe von Ruth Nussinov und Haim Wolfson sind eine ganze Reihe von Arbeiten zum Vergleich kompletter Rezeptorstrukturen oder Substrukturen entstanden [Bachar et al., 1993; Fischer et al., 1993a,b, 1994, 1995a,b; Lin et al., 1994; Lin and Nussinov, 1996; Norel et al., 1994; Rosen et al., 1998]. Die Proteininformation wird in diesen Ansätzen verschieden kodiert, zum Beispiel werden Cα-Atome, Oberflächenpunkte der Lösungsmittel zugänglichen Oberfläche nach Connolly (*Connolly surface*) oder sogenannte *sparse critical points*, eine reduzierte effiziente Beschreibung der Connolly Oberfläche, verwendet. Die Verfahren nutzen einen Geometric Hashing Algorithmus, um ähnliche Substrukturen zu entdecken und sind völlig unabhängig von der Sequenz oder Faltung der Proteine. Gerade der Ansatz von Rosen [Rosen et al., 1998] erlaubt einen automatisierten Vergleichsprozeß zweier Rezeptortaschen. Die Methoden wurden an einer ganzen Reihe von Proteinen validiert, unter anderem am Beispiel der Serinproteasen, den Chorismatmutasen, sowie Ähnlichkeiten im Faltungsmuster der Globinfamilien.

Einen ähnlichen Ansatz wie Fischer [Fischer et al., 1994] zum Auffinden von ähnlichen 3D Substrukturen in Protein verwenden Pennec et al. [Pennec and Ayache, 1998].

Erst kürzlich hat die Gruppe um R. Nussinov und H. Wolfson das Verfahren SiteEngine vorgestellt, daß Oberflächen von Proteinen vergleicht [Shulman-Peleg et al., 2004]. Sie verwenden das Pseudozentrum-Konzept von Cavbase, in dem sie die physikochemischen Eigenschaften durch 3D-Deskriptoren repräsentieren. SiteEngine vergleicht Proteinbindetaschen aber auch komplette Proteinstrukturen. Für die Repräsentation der Proteineigenschaften werden die Aminosäuren zuerst in Pseudozentren übersetzt und die Eigenschaft jedes Pseudozentrums auf die umliegenden Oberflächenpunkte projiziert. Dabei wird die Direktionalität der Wasserstoffbindung nicht berücksichtigt. Für jeden Oberflächenbereich wird der Schwerpunkt ('Oberflächenpseudozentrum') bestimmt und ein Deskriptor für die Krümmung der Oberfläche berechnet. Während des Vergleiches zweier Bindetaschen/Proteinstrukturen werden alle Triplets von drei Pseudozentren überlagert. In einem sehr schnellen Bewertungsschritt werden nur die Überlagerungen berücksichtigt, bei denen die Oberflächenschwerpunkte in der Überlagerung nahe

zusammen fallen, eine ähnliche physikochemische Umgebung und eine vergleichbare Oberflächenkrümmung aufweisen. Die 5000 besten Überlagerungen werden anschließend nach ihren RMSD-Abweichungen geclustert und die beste Überlagerung jedes Clusters in einem zweiten aufwendigeren Bewertungsschritt genauer untersucht. Dazu muß nicht nur der Oberflächenschwerpunkt, sondern jeder einzelne Oberflächenpunkt den Anforderungen des ersten Bewertungsschrittes entsprechen. Für die 100 besten Überlagerungen wird mit Hilfe eines Graphalgorithmus die maximale eins-zu-eins Zuordnung von Pseudozentren zweier Bindetaschen gefunden, deren RMSD-Abweichung und den Überlappungsgrad von hydrophoben und aromatischen Oberflächenbereichen berücksichtigt. An verschiedenen Szenarien ist die Methode validiert worden, z. B. Erkennung von Adenin-Bindestellen, Estradiol-Bindestellen oder bei der Funktionsannotierung eines unbekannten Proteins. Ein Vorteil der Methode ist, daß man sowohl mit kompletten Proteinstrukturen als auch mit Bindetaschen Datenbanksuchen durchführen kann. SiteEngine benutzt aber keine Methode zur automatischen Detektion von Bindestellen, diese werden über die Präsenz eines zuvor definierten Liganden zugewiesen (Proteinatome im 4 Å Radius um die Ligandenoberflächen).

Kinoshita et al. haben eine Datenbank von Proteinoberflächen (eF-site) aufgebaut, die zur Ähnlichkeitssuche benutzt werden kann [Kinoshita and Nakamura, 2003; Kinoshita et al., 2002]. Für jeden Punkt auf der Oberfläche eines Proteins wird das elektrostatische Potential, ein Hydrophobizitätswert und geometrische Attribute abgespeichert. Ein Cliquealgorithmus sucht nach ähnlichen Bereichen von Oberflächen, an denen die Oberflächenpunkte gleiche Eigenschaften aufweisen. Bindetaschen werden als Bereich im Radius von 5 Å um einen Liganden definiert. Ähnlichkeitssuchen wurden anhand von sechs Beispielen wie u.a. der katalytischen Triade der Serinproteasen, katalytischen Antikörpern mit Esterase-Eigenschaften und der Phosphat-Bindestellen von Nukleotiden, durchgeführt.

Eine Methode, die einen genetischen Algorithmus zur Suche struktureller Ähnlichkeiten in Rezeptorstrukturen verwendet, wurde von Lehtonen et al. mit dem Programm GENFIT vorgestellt [Lehtonen et al., 1999]. Die Ähnlichkeiten zweier Strukturen können dabei sowohl auf Ebene kompletter Faltungsmuster sowie auch auf Ebene lokaler Muster detektiert werden. Dabei wird zunächst eine Auswahl zufälliger Rotationsmatrizen und Translationsvektoren generiert, anhand derer zwei zu vergleichende Strukturen mehrfach überlagert werden. Jede dieser Überlagerungen wird analysiert und bewertet, wobei die gut bewerteten Überlagerungen eine Generation im genetischen Algorithmus 'überleben'. Die dazugehörigen Rotationen und Translationen werden dann weiter 'mu-

tiert', so daß schließlich die Strategie zu besseren Überlagerungen konvergiert. Durch verschiedene Vergleiche können mit diesem Verfahren strukturelle Ähnlichkeiten in Rezeptoren gefunden werden, die sich zum Teil über das gesamte Faltungsmuster erstrecken, aber auch lokale Substrukturmuster im Cα-Gerüst in der Nähe von Ligand- oder Metallbindebereichen erkennen lassen.

Poirrette et al. [Poirrette et al., 1997] benutzen ebenfalls einen genetischen Algorithmus (GA), um die Connolly Oberfläche von Proteinen zu vergleichen. Die Oberflächenpunkte werden mit geometrischen Informationen, z. B. Deskriptoren, die die Gestalt der Oberflächen widerspiegeln und einfachen Wasserstoffbrücken-Deskriptoren ausgestattet. Der GA wird zur Bewertung der erzeugten Rotationen und Translationen eingesetzt. Die Überlappung von Oberflächenpunkten dient als einfache Scoringfunktion. An sechs Testbeispielen (HIV Protease, Methotrexate Bindestelle (mit/ohne Ligand), Sialidasen verschiedener Spezies, NAD Bindestelle in Dehydrogenasen, Lysozym-Antikörper Komplex, Elastase Inhibitoren) wurde die Methode validiert.

Pickering et al. [Pickering et al., 2001] benutzen einen Clique Algorithmus, um Proteinoberflächen miteinander zu vergleichen. Als Bindestellen werden Bereiche in einem bestimmten Radius um den Liganden definiert. Connolly Oberflächenpunkte beschreiben die Proteinoberfläche und werden mit ihrer Kurvigkeit und einem Maß für ihre Gestalt ausgestattet. Die Methode wurde nur an einem Beispiel validiert, den NAD Bindetaschen aus verschiedenen Spezies.

Jambon et. al [Jambon et al., 2003] vergleichen Proteine ohne auf Sequenz- und Faltungsinformationen zurückzugreifen. Aminosäuren werden durch einen Satz funktioneller chemischer Gruppen repräsentiert, hydrophobe Aminosäuren werden nicht berücksichtigt. Ein Protein wird als Graph aus Dreiecken, die jeweils chemische Gruppen repräsentieren, dargestellt. Eine heuristische Suche erlaubt das Auffinden von ähnlichen Substrukturen in zwei Proteinen. An zwei knappen Beispielen (Serinproteasen, Proteine der Lectin Familie) wird die Methode validiert.

Stahl et al. [Stahl et al., 2000] berichteten über die automatische Gruppierung von 176 Zinkmetalloproteinasen mit Hilfe eines selbstorganisierenden Neuronalen Netzes. Dazu wurden Oberflächen-zugängliche Muster physikochemischer Eigenschaften in Bindetaschen ausgewertet. Der Ansatz separiert die katalytischen Zentren dieser Enzyme klar von anderen Kavitäten auf deren Oberfläche.

Methoden zur Detektion von Substrukturen gemeinsamer Oberflächenbereiche von Proteinen sind in der Lage, verwandte Proteine zu detektieren. Keiner der vorgestellten Algorithmen koppelt aber ein Verfahren zur automatischen Detektion von Bindestellen mit einem Vergleichs- und Suchalgorithmus. Gerade diese Kombination ist in Cavbase effizient implementiert und erlaubt automatische und effiziente Suchen nach funktionell bedeutsamen Ähnlichkeiten in großen Datensätzen.

Tab. 2.1 Auswahl an Methoden, die Proteine anhand ihrer 3D-Struktur vergleichen.

Referenz	Name	Ziel	Modellierung	Algorithmus	No.	GP[1]	Beispiele.
[Artymiuk et al., 1994]	ASSAM	Suche nach 3D-Mustern in Proteinen	2 Pseudoatome pro Aminosäure	Ullmann Subgraph Isomorphismus	4	nein	Serinproteasen, Zink-Bindung in Thermolysin, artifizielle Muster aus Aminosäuren
[Spriggs et al., 2003]	ASSAM	Suche nach Aminosäuremustern	2 Pseudoatome pro Aminosäure, zusätzlich Berücksichtigung der Hauptkettenatome durch drei Vektoren	Ullmann Subgraph Isomorphismus	1	nein	Serinproteasen
[Hamelryck, 2003]	—	Suche nach Mustern aus 3 Seitenketten	Zentralatom beschreibt Funktionalität der Aminosäure, nicht alle Atome der Aminosäuren werden benutzt. Vektoren beschreiben den räumlichen Aufbau	Nächste Nachbarnsuchen in multidimensionalen Indexbäumen (SR)	6	nein	Serinproteasen, spiegelbildliche Geometrien von Aminosäuremustern, umfassende Suche nach Mustern aus Triaden (fünf Beispiele) Elastase, etc.
[Wangikar et al., 2003]	DRESPAT	Detektion von Seitenkettenmustern unter Verwendung eines automatisierten Graph-basierten Ansatzes	siehe [Russell, 1998a]	Clique Algorithmus	17	ja	Ähnlichkeitssuche in SCOP-Datenbanken
[Pickering et al., 2001]	BALSAMIC	Beschreibung und Vergleich von Proteinoberflächen	Connolly Oberflächenpunkte beschrieben durch Kurvigkeit und Form (shape index)	Clique Algorithmus	1	beides	NAD Bindetaschen
[Kleywegt, 1999]	SPASM RIGOR	Suche mit einem Protein in einer Datenbank von vordefinierten Mustern. Suche in einer Datenbank von Proteinen mit einem Motiv als Anfrage	Cα-Atom und Pseudoatom am Schwerpunkt der Seitenkette beschreiben Aminosäure	Fuzzy Matching	4	nein	Retinol-bindendes Protein, Cellobiohydrolase, Thermolysin, Identifizierung von links-gängigen Helices
[Kinoshita et al., 2002] und [Kinoshita and Nakamura, 2003]	eF-site	Vergleich von Proteinoberflächen	Die Connolly Oberfläche wird mit dem elektrostatischen Potential, der Hydrophobizität der Seitenkette und der Flexibilität ausgestattet	Clique Algorithmus	6	beides	Serinproteasen, Phosphatbindestelle von Nukleotiden, katalytische Antikörper, PEP-Carboxy Kinase, hypothetisches Protein MJ0226
[Wallace et al., 1997]	TESS	Eine Datenbank aus 3D Templaten kann nach einem vorher definierten Muster durchsucht werden	Jede Aminosäure wird durch drei Punkte definiert, die in einem Hash-Tabelle abgelegt werden.	Geometric Hashing	3	nein	Serinproteasen, Ribonukleasen, Lysozym
[Jambon et al., 2003]	SUMO	Detektion von ähnlichen 3D-Stellen in Proteinen	Aminosäuren sind durch chemische Gruppen repräsentiert, die noch Informationen über geometrische und die lokale Umgebung der Gruppen (Dichte) enthalten	Graph Matching mit Dreiecken	2	ja	Serinroteasen, Lectine
[Lehtonen et al., 1999]	GENFIT	Detektion von gemeinsamen Substrukturen in Proteinen	Cα-Atom Koordinaten	Genetischer Algorithmus	9	beides	Serinproteasen, Cu/Ca-bindende Proteine,PH/OB-fold,ATP-Bindestellen, etc.

(Fortsetzung nächste Seite)

Fortsetzung Tab. 2.1

Referenz	Name	Ziel	Modellierung	Algorithmus	No.	GP[1]	Beispiele.
[Poirrette et al., 1997]	—	Vergleich der Solvent-zugänglichen Oberfläche zweier Proteine	Connolly Oberflächenpunkte ausgestattet mit Informationen über Form der Oberfläche, Wasserstoffbrücken-Deskriptoren und Oberflächennormalen	Genetischer Algorithmus	7	beides	HIV Protease, Sialidase, NAD, Elastase, Lysozym-Antikörper
[Pennec and Ayache, 1998]	—	3D-Substruktur Vergleiche	Aminosäuren werden durch drei Hauptkettenatome repräsentiert	Geometric Hashing	1	-	Tryptophan Repressor
[Fischer et al., 1993a,b]	—	Strukturelle (Cα-Atom) und Oberflächenvergleiche von Proteine	Cα-Atome, Connolly- und SPHGEN-Oberflächenpunkte	Geometric Hashing	2	-	Hämoglobin/Myoglobin und ADH/LDH
[Fischer et al., 1994]	—	Detektion von Motiven in Proteinen	Cα-Atome	Geometric Hashing	2		Serinproteasen, Sulfhydryl Proteasen
[Rosen et al., 1998]	—	Vergleich von molekularen Oberflächen zur Identifizierung von Bindetaschen und zum Auffinden von Ähnlichkeiten zwischen Proteinen	Deskriptoren (sparse critical points), die die Eigenschaften der Connolly Oberfläche beschreiben	Geometric Hashing	2	beides	Serinproteasen, Chorismate Mutasen
[Schmitt et al., 2001]	CAVBASE	Beschreibung und Vergleich von Proteinbindetaschen	Physikochemische Interaktionsmöglichkeiten der Aminosäuren	Clique Algorithmus	6	nein	Serinproteasen, Phosphatbindestellen, Adeninbinderegionen, Chorismatmutasen, NADPH-Bindestellen
[Russell, 1998a] und [Stark et al., 2003]	PINTS	Detektion von Seitenkettenmustern	Aminosäuren werden durch einzelnes funktionelles Atom repräsentiert, Räumliche Anordnung zwischen Aminosäuren wird durch Distanzen zwischen Cα-, Cβ-Atomen wiedergegeben.	*Depth-First* Suchalgorithmus	8	nein	Serinproteasen, Eisen-bindende Protein, Zinkfinger-Proteine, SCOP Repräsentative, Phosphatase/Aminopeptidase, DNase/Endocellulase E1, Chitobiase/Neuraminidase
[Liang et al., 2003b]	web FEATURE	Vorhersage von Bindestellen durch Analyse von 3D Motiven, die von Sequenzmotiven abgeleitet werden	Verschiedene Atom-basierte physikochemische Deskriptoren	Statistische Analyse, *Bayesian inference method*	1	nein	EF-Hand Proteine (Calcium Bindung)
[Laskowski et al., 2003]	ProFunc	Vorhersage der Funktion von Proteine unter Verwendung verschiedener Ansätze und Tools	verschiedene 2D und 3D Methode	Jess	3	ja	Hypothetische Proteine wie YbgJ, CutA, MT777
[Stahl et al., 2000]	—	Klassifizierung von Zink-bindenden Proteinen	Connolly Oberflächenpunkte, die mit fünf generischen Eigenschaften ausgestattet ist	neuronales Netz	1	nein	176 Zink-bindende Proteine im Trainingsdatensatz und 18 im Testdatensatz
[de Rinaldis et al., 1998]	3D-Profile	Beschreibung und Vergleich von Proteinoberflächen	3D Profile aus multiplen Sequenzalignments; Aminosäuren werden durch ein Atom der Seitenkette wiedergegeben	Erschöpfende Suche	3	ja	Proteine mit SH3- und SH2-Domänen, Proteine mit P-Bindeschleife

[1] ganzes Protein (GP): Ist es möglich mit diesem Ansatz gesamte Proteinstrukturen zu vergleichen?

3 Theorie und Methoden

In diesem Kapitel wird die Methode Cavbase kurz vorgestellt und auf Erweiterungen und Optimierungen eingegangen, die im Rahmen dieser Arbeit durchgeführt wurden. Ein Ziel dieser Arbeit war, zwei Aspekte an Cavbase zu optimieren: Einmal sollte eine Validierung der Parameter, die bei der Generierung und Charakterisierung einer Bindetasche eine Rolle spielen so optimiert werden, daß die physikochemischen Eigenschaften möglichst vollständig repräsentiert werden. Zum anderen sollte die Geschwindigkeit der Bindetaschenvergleiche signifikant erhöht werden, um den Vergleich von großen Datensätze zu ermöglichen. Ein kurzer Überblick über Cavbase wird in Abschnitt 5.1 gegeben, in Kapitel 3.2 wird auf die Repräsentierung der Aminosäuren und ihrer Eigenschaften eingegangen. In Abschnitt 3.3 wird eine Beschreibung der Bindetaschenoberfläche durch Oberflächenpunkte mit multiplen Eigenschaften vorgestellt und in Abschnitt 3.4 wird eine qualitative Validierung dieser Eigenschaften diskutiert. Auf die Optimierung und Beschleunigung der Bindetaschenvergleiche wird in Abschnitt 3.5 eingegangen. Das Kapitel schließt mit einer Darstellung der Entwicklung von Bindetaschenbitstrings und deren Einsatz für den Vergleich von Proteinbindetaschen (Abschnitt 3.6).

3.1 Cavbase - eine Methode zum Beschreiben und zum Vergleich von Proteinbindetaschen

In diesem Abschnitt wird ein kurzer Überblick über die Methode Cavbase gegeben. In einer vorherigen Arbeit wurde Cavbase als eine Methode zur Beschreibung und Vergleich von Proteinbindetaschen entwickelt [Schmitt, 2000; Schmitt et al., 2001, 2002]. Cavbase ist eine Datenbank von Proteinbindetaschen, die Informationen über die Form und die physikochemischen Eigenschaften von Bindetaschen von allen bekannten Proteinstrukturen enthält[1].

Cavbase detektiert in einem automatisierten Verfahren Vertiefungen auf der Proteinoberfläche als mögliche Bindestellen (siehe Abschnitt 3.3). Dabei wird keine Information über eventuell gebundene Liganden verwendet, sondern eine Bindetasche wird nach rein

[1] Cavbase enthält die Information über 80661 Bindetaschen, die aus 22885 Proteinen ausgeschnitten wurden (Stand Juli 2004).

geometrischen Gesichtspunkten definiert. Eine Bindetasche besteht aus Aminosäuren, einer Bindetaschenoberfläche und eventuell gebundenen Liganden. Die physikochemischen Eigenschaften der Aminosäuren werden durch 3D-Deskriptoren (Pseudozentren) kodiert, die wichtige Eigenschaften der molekularen Erkennung repräsentieren (siehe Abschnitt 3.2): Wasserstoffbrücken-Bindungspartnereigenschaften, hydrophobe und aromatische Wechselwirkungen. Die Eigenschaft der Pseudozentren wird auf die Bindetaschenoberfläche projiziert und die Oberfläche mit der entsprechenden Eigenschaft annotiert. Dabei wird darauf geachtet, daß nur die Pseudozentren berücksichtigt werden, die ihre Eigenschaft auf die Oberfläche projizieren und so mit einem potentiellen Liganden interagieren können (siehe Abschnitt 3.3). In Abbildung 5.1 ist die Detektion der Bindetaschen und die Zuweisung der Pseudozentren noch einmal bildlich dargestellt.

Bindetaschen werden mit einem Clique-Algorithmus verglichen (siehe Abschnitt 3.5). Eine Bindetasche kann als Graph mit den Pseudozentren als Knoten aufgefasst werden. Der Clique-Algorithmus ist in der Lage, nach gemeinsamen Mustern von Pseudozentren in zwei Bindetaschen zu suchen, um so Ähnlichkeiten zu entdecken. Anschließend berechnet ein Scoringschritt (siehe Abschnitt 3.5), wie gut die Oberflächen beider Taschen überlappen. Dadurch ist es möglich, die Ähnlichkeit zwischen zwei Bindetaschen zu quantifizieren.

3.2 Validierung der Aminosäurenrepräsentation in Cavbase

In Cavbase werden die physikochemischen Eigenschaften der Aminosäuren durch 3D-Deskriptoren (Pseudozentren) kodiert, die wichtige Eigenschaften der molekularen Erkennung repräsentieren: Wasserstoffbrücken-Donoren, -Akzeptoren, ambivalente -Donoren/-Akzeptoren (z.B. Hydroxylgruppen oder Seitenkettenstickstoffe des Histidins), hydrophobe aliphatische und aromatische Eigenschaften (siehe auch Abschnitt 5.1). In Abbildung 3.2 sind die Pseudozentren gezeigt, die in einer vorherigen Arbeit entwickelt und verwendet wurden [Schmitt, 2000]. Alle Aminosäuren, die an der Ausbildung einer Bindetasche beteiligt sind, werden in Pseudozentren übersetzt. Die Repräsentation der physikochemischen Eigenschaften einer Bindetasche durch Pseudozentren stellt eine effiziente Form der Beschreibung dar. Zum einen werden die Aminosäuren durch einen reduzierten Satz von 3D-Deskriptoren repräsentiert und

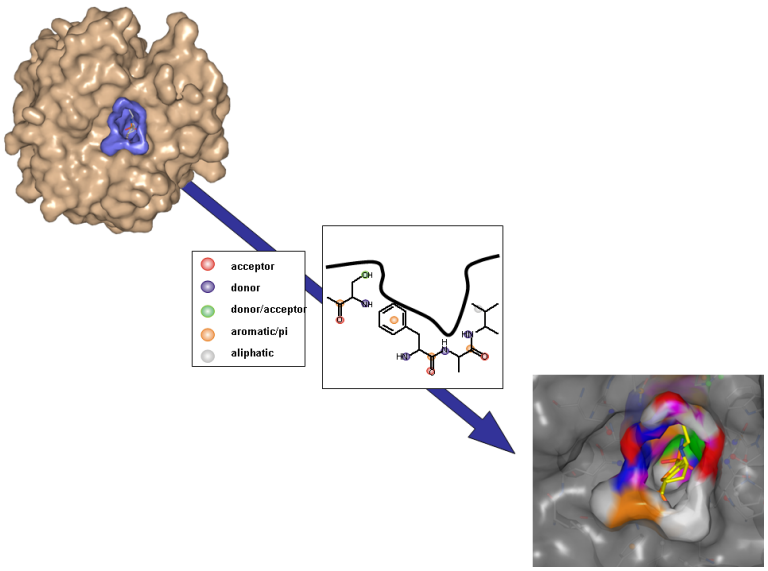


Abb. 3.1 Von der Proteinstruktur zur Bindetasche. Bindetaschen werden als Vertiefungen (blaue Oberfläche) auf der Oberfläche von Proteinen detektiert. Die physikochemischen Eigenschaften der Aminosäuren werden durch 3D Deskriptoren (Pseudozentren) wiedergegeben. Sie werden als farbige Kugeln dargestellt: Wasserstoffbrücken-Donor (blau), -Akzeptor (rot), -Donor/Akzeptor (grün), hydrophobe Wechselwirkungen (weiß) und aromatische Wechselwirkungen (orange). Es werden nur die Pseudozentren zur Beschreibung der Bindetaschen berücksichtigt, die ihre Eigenschaft auf die Oberfläche exponieren können. Eine Bindetasche, wie sie in Cavbase abgelegt ist, besteht aus Aminosäuren, Pseudozentren, einer Bindetaschenoberfläche und eventuell gebundenen Liganden.

dadurch ein gewißes Maß an Abstraktion von den tatsächlich vorhandenen Aminosäuren hin zu den wichtigen physikochemischen Eigenschaften erreicht. Zum anderen wird die Komplexität in Ähnlichkeitsanalysen wesentlich gesenkt, da deutlich weniger 3D-Koordinaten betrachtet werden müssen.

Datenbanken wie Isostar [Bruno et al., 1997] oder HB-Atlas [McDonald and Thornton, 1994] enthalten Informationen über in Kristallstrukturen beobachtete Wechselwirkungsgeometrien zwischen Aminosäuren. Wird eine Interaktion zwischen zwei bestimmten Gruppen in großer Zahl beobachtet, kann davon ausgegangen werden, daß die Ausbildung dieser Interaktion unter energetischen Gesichtspunkten möglich bzw. günstig ist.

Die Isostar-Datenbank enthält Informationen über Wechselwirkungsgeometrien von Fragmenten bzw. funktionellen Gruppen mit Nachbargruppen, wie sie in der Kristallpackung beobachtet werden. Die statistische Verteilung der Fragmente wird als Scatterplot dargestellt, sie zeigt die Verteilung einer bestimmten funktionellen Gruppe als

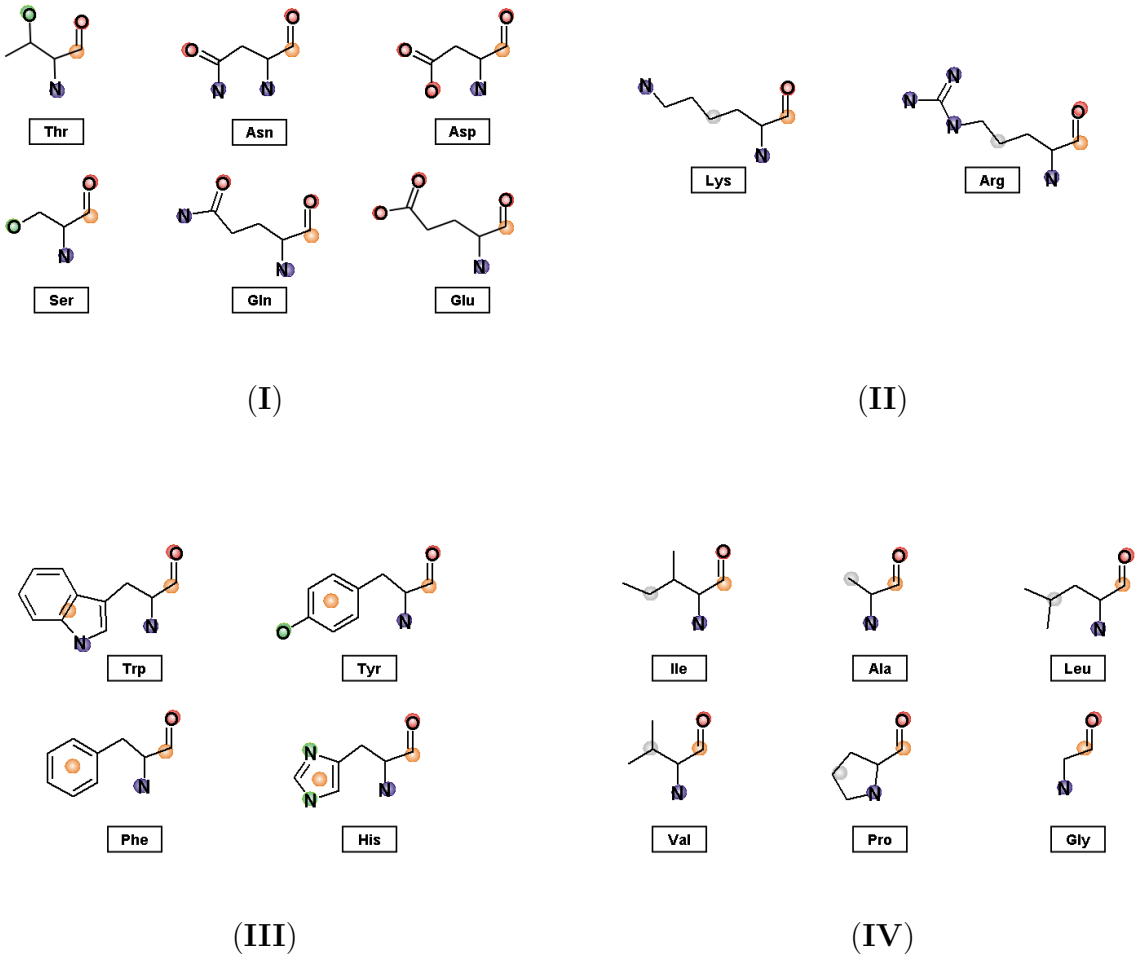


Abb. 3.2 Repräsentation der Aminosäuren in Cavbase nach [Schmitt, 2000]. In I und II sind polare Aminosäuren, in (III) Aminosäuren mit aromatischem Charakter und in (IV) mit aliphatischen Eigenschaften dargestellt. Die Position der Pseudozentren wird durch farbige Kugeln angedeutet. Farbcodierung: Wasserstoffbrücken-Donor (blau), Wasserstoffbrücken-Akzeptor (rot), Wasserstoffbrücken-Donor und -Akzeptor (grün), aromatische Wechselwirkungen (orange), hydrophobe Wechselwirkungen (weiß). Zu beachten ist, daß die Position der hydrophoben Pseudozentren dem geometrischen Schwerpunkt der hydrophoben Atome der jeweiligen Seitenketten entspricht. Dabei werden die Beiträge der einzelnen Atomkoordinaten noch entsprechend ihrer Fähigkeit, wie gut sie ihre Eigenschaft auf die benachbarten Punkte der Bindetaschenoberfläche exponieren können, gewichtet.

Sonde um ein bestimmtes Strukturfragment als zentrale Gruppe. Um die Interaktionsmöglichkeiten der Aminosäuren besser beurteilen zu können, wurden die Scatterplots für die Peptidbindung und die Seitenketten der Aminosäuren untersucht. Die Datenbank HB-Atlas enthält Informationen über das Wasserstoffbrücken-Bindungsverhalten

von Aminosäuren. Die Datenbank wurde durch die Analyse von Kristallstrukturen manuell erstellt und die Häufigkeit, mit der bestimmte Aminosäuren Wasserstoffbrücken ausbilden, bestimmt. Beide Datenbanken erlauben es, systematisch Interaktionen zwischen Aminosäurefragmenten zu untersuchen und helfen so bei der Zuweisung neuer Pseudozentren.

Die Analyse der Wasserstoffbrücken-Muster, an denen Cystein beteiligt ist, hat gezeigt, daß die Sulfhydrylgruppe als Partner von Wasserstoffbrücken-Wechselwirkungen auftritt. Diese Annahme wird noch dadurch unterstützt, daß Cystein eine wichtige Rolle in katalysierten Reaktionen spielt und oft als katalytisch essentieller Rest in Bindetaschen zu finden ist [Bartlett et al., 2002]. In der Analyse von McDonald und Kollegen [McDonald and Thornton, 1994] hat sich gezeigt, daß Cystein sehr selten als Wasserstoffbrücken-Akzeptor fungiert, sehr wohl aber als Wasserstoffbrücken-Donor. Die Länge der Wasserstoffbindung ist im Vergleich größer als eine NH-Sauerstoff Wasserstoffbrücke, da das Schwefelatom einen größeren Atomradius als ein Sauerstoffatom hat. Eine visuelle Analyse unterschiedlicher Verteilungen von verschiedenen Wasserstoffbrücken-Donorgruppen um Schwefel-haltige Aminosäuren zeigt (siehe Abbildung 3.3), daß sich Kontaktgruppen mit keiner eindeutig bevorzugten Geometrie unter Ausbildung einer Wasserstoffbrücke um die Thiogruppe scharen. Deshalb wurde nur ein Donor-Pseudozentrum, lokalisiert am Schwefelatom des Cysteins, eingefügt. Der aliphatische Charakter des Cysteins wird weiterhin durch ein aliphatisches Pseudozentrum, lokalisiert auf der Seitenkette, ausgedrückt. Der Schwefel in der Seitenkette des Methionins zeigt keine deutlich hervorgehobenen Wasserstoffbrücken-Bindungseigenschaften, daher wird diese Seitenkette durch ein aliphatisches Pseudozentrum repräsentiert.

Untersuchungen zur Wechselwirkung von Aminosäuren mit aromatischen Systemen [Mitchell et al., 1994; Meyer et al., 2003] sowie Analysen mit Isostar haben gezeigt, daß die Seitenketten von Asparagin, Glutamin, Aspartat, Glutamat und Arginin π-Wechselwirkungen mit benachbarten Liganden ausbilden können. Analysiert man beispielsweise die Verteilungen von aromatischen Kontaktgruppen um Säureamidgruppen als Zentralfragment (siehe Abbildung 3.3-IV), ist deutlich die Präferenz bestimmte Geometrien zu beobachten. Um diese Interaktionsmöglichkeiten der Aminosäuren wiedergeben zu können, wurde auf dem Kohlenstoffatom einer Carboxylat-, Säureamid- bzw. Guanidin-Gruppe ein Pi Pseudozentrum eingeführt. Die Fähigkeit von Carboxylat-Gruppen zur Ausbildung von $\pi - \pi$-Wechselwirkungen wird im Vergleich zu Säureamidgruppen als schwächer eingeschätzt [Mitchell et al., 1994]. Analysen von Kristall-

strukturdaten haben allerdings gezeigt, daß diese Wechselwirkungen durchaus beobachtet werden. Zusätzlich sollen in einer Ähnlichkeitsanalyse zwei der drei Seitenkettenpseudozentren von beispielsweise Glutamin und Glutamat als ähnlich erkannt werden, weshalb für Aspartat und Glutamat ebenfalls ein entsprechendes Pi Pseudozentrum eingeführt wurde.

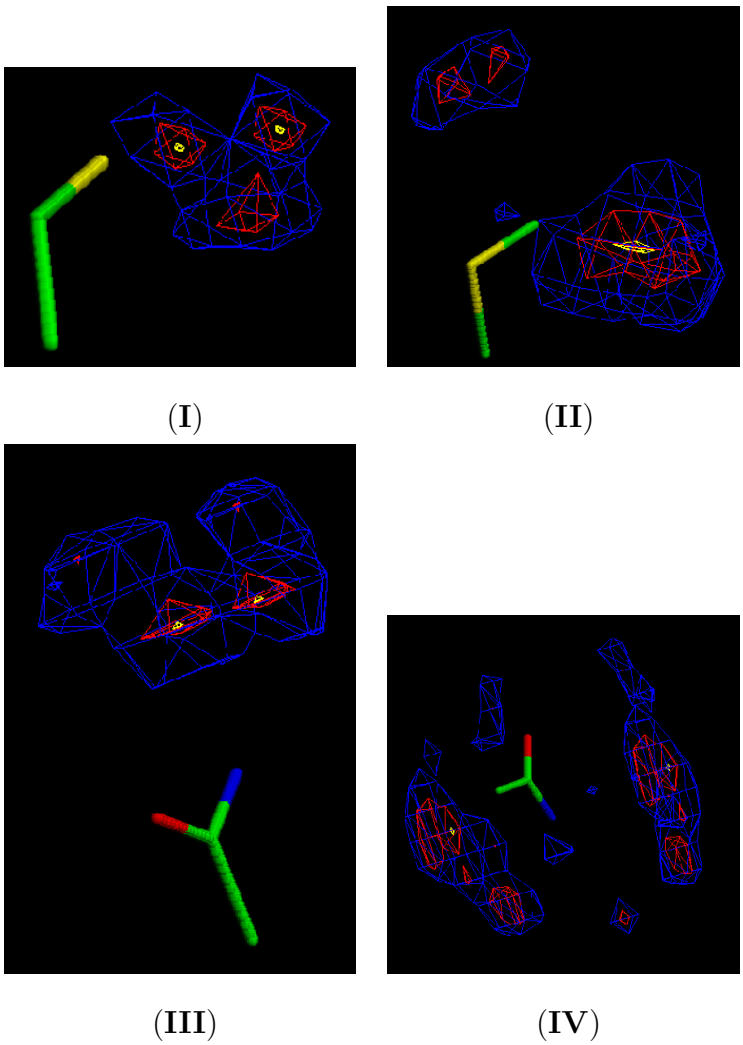


Abb. 3.3 Experimentell beobachtete Verteilungen einer Carbonyl-Kontaktgruppe um Schwefel-haltige Aminosäuren und aliphatischen und aromatischen CH-Gruppen um eine Säureamidgruppe. (I) und (II) zeigen die Verteilung eines Carbonyl-Sauerstoffs (*any carbonyl*) um eine terminale Thiol-Gruppe (Cystein) und um eine Methylthioether-Gruppe (Methionin). Die Interaktion zwischen einem Cystein und dem Carbonylsauerstoff ist gerichteter und räumlich stärker konzentriert. Die Isostar Analyse unterstützt die Vermutung, daß Cystein - aber nicht Methionin - in der Lage ist, als Wasserstoffbrückenpartner aufzutreten. (III) und (IV) zeigen die Verteilung von aliphatischen CH-Gruppen und aromatischen CH-Gruppen um eine Säureamidgruppe (Asparagin, Glutamin). Die Analyse der Interaktion der Säureamidgruppe mit aliphatischen CH-Gruppen zeigt keine Präferenzen für die Ausbildung bestimmter Geometrien (III), die auf eine ausgeprägte Wechselwirkung schließen lässt. In (IV) sind Interaktionen von Kohlenstoffen aus aromatischen Ringsystemen und der Säureamidgruppe zu erkennen, die gehäuft oberhalb und unterhalb der Ebene, die durch die Atome der Säureamidgruppe aufgespannt werden, auftreten. Eine Analyse der Geometrien, die die aromatischen Ringsysteme zu der Säureamidgruppe einnehmen, zeigt neben *face-to-face*-artigen auch *edge-to-face*-artige Geometrien. Es läßt sich hier keine eindeutige Präferenz für eine dieser Geometrien erkennen (siehe auch Abschnitt 3.3.2). Allerdings werden diese Interaktionen gehäuft oberhalb und unterhalb der Ebene, die durch die Atome der Säureamidgruppe aufgespannt werden, beobachtet.

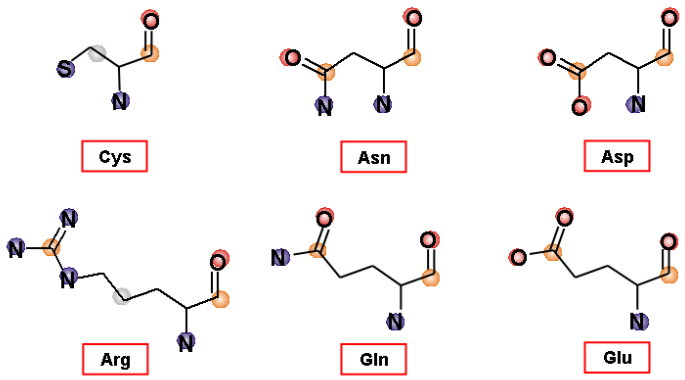


Abb. 3.4 In Erweiterung zur ursprünglichen Version von Cavbase neu zugewiesene Pseudozentren. Die Fähigkeit der Thiolgruppe des Cysteins zur Ausbildung von Wasserstoffbrücken wird durch ein Donor Pseudozentrum repräsentiert. Den Seitenketten von Asparagin und Glutamin, Aspartat und Glutamat und Arginin wird ein Pi Pseudozentrum zugewiesen, um ihrer Fähigkeit π-π-Wechselwirkungen ausbilden zu können, Rechnung zu tragen.

3.3 Oberflächenpunkte mit multiplen Eigenschaften erlauben eine unscharfe Modellierung der Bindetasche

3.3.1 Oberflächenpunkte mit multiplen Eigenschaften[2]

Im ursprünglichen Cavbase Ansatz werden Bindetaschen mit dem Ligsite Algorithmus detektiert [Hendlich et al., 1997] und die Oberflächenpunkte bekommen genau eine Eigenschaft zugewiesen [Schmitt, 2000]. In diesem Abschnitt wird die Entwicklung von Oberflächenpunkten mit multiplen Eigenschaften vorgestellt. Zuerst soll ein kurzer Überblick über den Bindetaschendetektionsalgorithmus und die Zuweisung der Bindetaschenoberfläche gegeben werden:

- Das Protein wird in ein Gitter eingebettet, die Gitterweite beträgt 0.5Å.
- Alle Gitterpunkte, die innerhalb des van der Waals (vdW) Radius um Proteinatome zu liegen kommen, werden in der Analyse nicht weiter berücksichtigt.
- Die verbleibenden Gitterpunkte werden nach der Vergrabenheit im Protein bewertet, d.h. ob sie sich beispielsweise an der Proteinoberfläche (niedrige Vergrabenheit) oder tief im Protein in einer Bindetasche (hohe Vergrabenheit) befinden.
- Der Vergrabenheitswert eines Gitterpunktes im Protein kann Werte von Null bis Sieben annehmen. Es wird überprüft, ob der Gitterpunkt von Proteinatomen eingeschlossen ist. Dabei wird von dem Gitterpunkt ausgehend die Bindetasche entlang der drei Hauptkoordinatenachsen (x,y,z) und der vier Raumdiagonalen abgetastet. Wenn ein Gitterpunkt entlang einer Koordinatenachse von beiden Seiten von Rezeptoratomen eingeschlossen ist, wird der Vergrabenheitswert des entsprechenden Gitterpunktes um Eins erhöht. Somit kann der maximale Vergrabenheitswert für Gitterpunkte, die von allen Seiten von Proteinatomen eingeschlossen sind, sieben betragen.
- Alle Gitterpunkte, die nicht Solvent-zugänglich sind, werden in der Bestimmung des Vergrabenheitswertes nicht berücksichtigt. Zur Bestimmung der Solvent-

[2] *Die Arbeiten an der Repräsentation der Bindetaschenoberfläche mit multiplen Eigenschaften, die in diesem Abschnitt vorgestellt werden, wurden in Zusammenarbeit mit Nils Weskamp (Universität Marburg) durchgeführt.*

zugänglichen Gitterpunkte wird die Proteinoberfläche mit einer Probensonde (Radius 1.5Å) abgetastet.

- Nachdem alle Gitterpunkte einen Vergrabenheitswert zugewiesen bekommen haben oder in der weiteren Analyse nicht mehr berücksichtigt werden, werden benachbarte Gitterpunkte, die einen bestimmten Vergrabenheitswert besitzen (beispielsweise > 3), zu Clustern zusammengefasst. Die so gebildeten Cluster von Gitterpunkten stellen die Bindetasche in ihrer gesamten Form dar.

- Die Randpunkte eines gefundenen Clusters approximieren die Bindetaschenoberfläche.

- Jede Aminosäure, deren Atome sich in einer bestimmten Distanz zu einem Oberflächenpunkt befinden, wird zur Bindetasche gezählt. Die Distanz berechnet sich aus der Summe des vdW Radius des Atoms plus einem vorgegebenem Wert von 1.1Å.

- Nach einem festen Regelsatz werden die Aminosäuren, die die Bindetasche begrenzen, in Pseudozentren übersetzt ([Schmitt, 2000], Abschnitt 3.2).

- Nach der Übersetzung der Aminosäuren in einen Satz von Pseudozentren werden diese in einem Filterschritt daraufhin untersucht, ob sie ihre Eigenschaft auf die Bindetaschenoberfläche ausrichten.

  - Für alle Pseudozentren, die sich innerhalb der Interaktionsdistanz zur Bindetaschenoberfläche befinden (3.0Å für nicht-aliphatische Pseudozentren und 3.5Å für aliphatische Pseudozentren) werden zwei Vektoren berechnet. Ein Standard-Vektor $\vec{v}$ beschreibt die Raumrichtung, in der die jeweilige Eigenschaft im Mittel exponiert wird. Ein zweiter Richtungsvektor $\vec{r}$ zeigt in Richtung Bindetaschenoberfläche. Dieser ergibt sich als normierter Summenvektor aus allen Ortsvektoren, die von dem Pseudozentrum zu den benachbarten Oberflächenpunkten ausgehen [Schmitt, 2000].
  - Pseudozentren werden in der weiteren Analyse nur dann berücksichtigt, wenn der eingeschlossene Winkel nicht größer als definierte *cut-off*-Werte[3] ist.

- Für die verbleibenden Pseudozentren wird nun geprüft, auf welche Oberflächenpunkte der Bindetaschenoberfläche sie ihre Eigenschaft abbilden können:

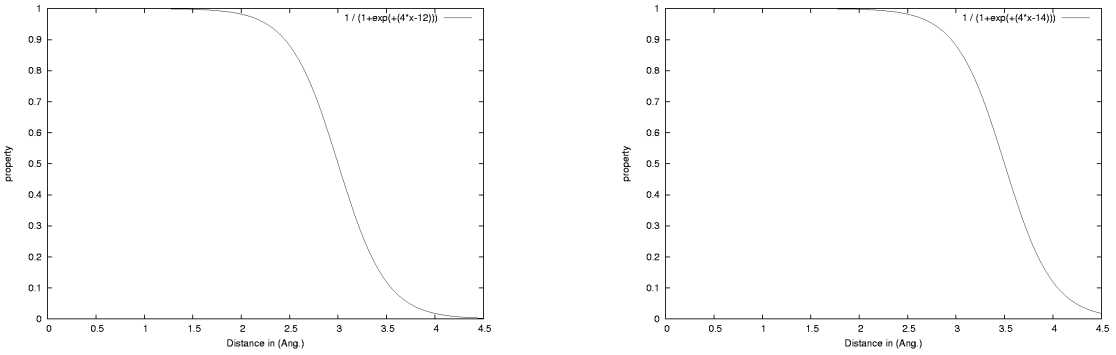
[3] Verwendete *cut-off*-Werte für $\measuredangle(\vec{v};\vec{r})$: Donor = 100^o; Akzeptor=,100^o, Donor-Akzeptor=100^o, Pi = 60^o

- Für jedes Pseudozentrum wird eine Schleife über alle Oberflächenpunkte durchlaufen und die Distanz zwischen Oberflächenpunkt und Pseudozentrum bestimmt. Ist sie kleiner als eine vorgegebene Distanz (3.0 Å für nicht-aliphatische Pseudozentren oder 3.5 Å für aliphatische Pseudozentren), wird dem Oberflächenpunkt die Eigenschaft des Pseudozentrums zugewiesen. Die Distanzen geben den Bereich wieder, in dem die jeweilige Wechselwirkung ausgebildet werden kann.
- Wenn ein Oberflächenpunkt bereits die Eigenschaft eines Pseudozentrums zugewiesen bekommen hat und ein weiteres Pseudozentrum in kürzerer Distanz gefunden wird, wird ihm dessen Eigenschaft zugewiesen.

Häufig befinden sich Oberflächenpunkte aber in einer Interaktionsdistanz zu mehreren Pseudozentren, so daß in diesem Raumsegment prinzipiell mehrere Arten von Wechselwirkungen ausgebildet werden können. Um dieses Verhalten modellieren zu können, wird jeder Oberflächenpunkt mit einem vier-dimensionalen Vektor ausgestattet: $\vec{V} = (\mathsf{Donor}, \mathsf{Akzeptor}, \mathsf{Aliphatisch}, \mathsf{Pi})$. Der Vektor ist aus vier Fließkommazahlen zusammengesetzt, die jeweils den Anteil der bestimmten physikochemischen Eigenschaft beschreiben. Jeder Oberflächenpunkt wird mit einem solchen Vektor ausgestattet und kann nun mehrere Eigenschaften in unterschiedlichem Ausmaß besitzen. Die Werte werden Distanz-abhängig über eine sigmoide Funktion (Gl. 3.1) zugewiesen. Die Funktionen für aliphatische und nicht-aliphatische Wechselwirkungen werden so adaptiert, daß sie in etwa die Reichweite der jeweiligen Wechselwirkung umfassen[4]. Die Funktion nimmt einen maximalen Wert von 1.0 bei einer Distanz von 1.3 Å an und weist mit größeren Abständen abfallende Werte zu. Bei einer Distanz von 3.0 Å (nicht-aliphatische Wechselwirkungen) bzw. 3.5 Å (aliphatische Wechselwirkungen) (siehe Abbildung. 3.5) wird jeweils ein Eigenschaftswert von 0.5 zugewiesen. Da die Ähnlichkeit von zwei Oberflächenbereichen über ein Skalarprodukt berechnet wird (siehe Gleichung 3.4), zeigen zwei Oberflächenbereiche, bei denen eine physikochemische Eigenschaft jeweils den Wert 0.5 besitzt, nur eine sehr geringe Ähnlichkeit. Um dieses Verhalten zu modellieren wurden für nicht-aliphatische Wechselwirkungen folgende Parameter für $t = 12$ und $s = 4$ gewählt, während für aliphatische Wechselwirkungen die Parameter $t = 14$ und $s = 4$ verwendet wurden.

$$f(x) = \frac{1}{(1 + \exp(1 + (s \cdot x - t)))} \tag{3.1}$$

[4]Die verwendeten Werte für eine aliphatische Wechselwirkung betragen 3.5 Å, für nicht aliphatische Wechselwirkungen 3.0 Å



(**I**) (**II**)

Abb. 3.5 Sigmoide Funktionen werden benutzt, um den Bindetaschenoberflächenpunkten Distanz-abhängig Eigenschaftswerte der naheliegenden Pseudozentren zuzuweisen. In (I) wird die Funktion für nicht aliphatische Wechselwirkungen, in (II) die Funktion für aliphatische Wechselwirkungen gezeigt. Die Funktion weist Werte zwischen Eins und Null zu und erreicht bei 3.0 bzw. 3.5 Å einen Wert von 0.5. Da die Ähnlichkeit zwischen zwei Oberflächenbereichen über die Bildung eines Skalarprodukts berechnet wird, gehen zwei Eigenschaften, die jeweils den Wert 0.5 besitzen, mit einem Ähnlichkeitsmaß von 0.25 in die Berechnung des Skalarproduktes ein. Zwei Oberflächenbereiche gelten dann als ähnlich, wenn das Skalarprodukt über 0.7 liegt.

Die SURFACE Klasse in Cavbase wurde mit einem zusätzlichen Attribut *Eigenschaftsvektor* ($\vec{V} = (Donor, Akzeptor, Aliphatisch, Pi)$*.)* ausgestattet. In dem ursprünglichen Ansatz wurden zwei Pseudozentren als ähnlich angesehen, wenn die zugrundeliegenden Pseudozentren einen vergleichbaren Typ aufweisen. Bei der Verwendung von Oberflächenpunkten mit multiplen Eigenschaften wird die Ähnlichkeit von Oberflächenbereichen durch die Berechnung des Skalarprodukts zweier Vektoren berechnet. Wenn der Skalarwert über einem festgelegtem Wert (0.7) liegt, gelten beiden Oberflächenbereiche als ähnlich. Dadurch wird die Modellierung der Oberflächen unschärfer gefasst, da sich die Eigenschaften nicht zu 1.0 aufsummieren müssen, um als ähnlich erkannt zu werden.

- Ein Oberflächenbereich (*Patch*) besteht aus einer Menge von Oberflächenpunkten, die alle mit einem Eigenschaftsvektor ausgestattet sind:

$$\vec{v}_1 = \begin{pmatrix} v_{11} \\ \vdots \\ v_{14} \end{pmatrix}, \ldots, \begin{pmatrix} v_1 \\ \vdots \\ v_4 \end{pmatrix}_n = \begin{pmatrix} v_{n1} \\ \vdots \\ v_{n4} \end{pmatrix}$$

- Der entsprechende Summenvektor $\vec{v}$ eines Oberflächenbereiches berechnet sich nach Gleichung 3.2.

$$\vec{v} := \begin{pmatrix} v_1 \\ \vdots \\ v_4 \end{pmatrix} := \sum_{i=1}^{n} \vec{v}_i := \begin{pmatrix} \sum_{i=1}^{n} v_{i1} \\ \vdots \\ \sum_{i=1}^{n} v_{i4} \end{pmatrix} \tag{3.2}$$

- Der Summenvektor wird normalisiert, indem jede Komponente des Vektors durch die Anzahl aller Oberflächenpunkte eines Oberflächenbereiches dividiert wird (Gl. 3.3).

$$\vec{v}_{norm} := \frac{1}{n} \cdot \vec{v} := \begin{pmatrix} \frac{\sum_{i=1}^{n} v_{i1}}{n} \\ \vdots \\ \frac{\sum_{i=1}^{n} v_{i4}}{n} \end{pmatrix} \tag{3.3}$$

- Jede Komponente des normalisierten Summenvektors $\vec{v}_{norm}$, deren Wert kleiner als 0.2 ist, wird auf Null gesetzt. Eine Eigenschaft eines Oberflächenbereichs wird dadurch nur dann berücksichtigt, wenn sie zumindest auf $\frac{1}{5}$ des Oberflächenbereiches exponiert wird. Dieser so erhaltene Vektor $\vec{v}_{patch}$ (im Folgenden *Patch Vektor* genannt) beschreibt die physikochemische Zusammensetzung eines Oberflächenbereichs.

- Die Ähnlichkeit (S_{AB}) zweier Oberflächenbereiche berechnet sich durch das Skalarprodukt der beiden *Patch Vektoren* $\vec{v}_{patch}$ und $\vec{w}_{patch}$. Zwei Oberflächenbereiche gelten dann als ähnlich, wenn das Skalarprodukt größer einem bestimmten Wert ($overlap_{cutoff} = 0.7$) ist (Gl. 3.4).

$$S_{AB} = \begin{cases} \langle \vec{v}_{patch}, \vec{w}_{patch} \rangle & \text{wenn} \quad \langle \vec{v}_{patch}, \vec{w}_{patch} \rangle \geq overlap_{cutoff} \\ 0 & \text{sonst} \end{cases} \tag{3.4}$$

Ein Donor-Akzeptor Typ wird nicht mehr ausdrücklich verwendet, sondern dadurch modelliert, indem man der Donor- und Akzeptor-Eigenschaft gleichzeitig das Produkt

aus einem Skalierungsfaktor f und dem Distanz-abhängigen Eigenschaftswert zuweist. Der Skalierungsfaktor wird auf 0.7 gesetzt, damit der Vergleich von zwei Donor-Akzeptor *Patch Vektoren* $(\vec{v_1}, \vec{v_2})$ eine Ähnlichkeit von ungefähr Eins ergibt.

$$\begin{aligned} \vec{v_1} &= (0.7, 0.7, 0.0, 0.0) \\ \vec{v_2} &= (0.7, 0.7, 0.0, 0.0) \\ S_{AB} &= 0.7 \cdot 0.7 + 0.7 \cdot 0.7 + 0.0 \cdot 0.0 + 0,0 \cdot 0.0 \quad = \quad 0.98 \end{aligned}$$

Die Ähnlichkeit zwischen einem Donor-Akzeptor Oberflächenbereich $\vec{v_1}$ und einem reinem Akzeptor Oberflächenbereich $\vec{v_2}$ berechnet sich nach:

$$\begin{aligned} \vec{v_1} &= (0.7, 0.7, 0.0, 0.0) \\ \vec{v_2} &= (0.0, 1.0, 0.0, 0.0) \\ S_{AB} &= 0.7 \cdot 0.0 + 0.7 \cdot 1.0 + 0.0 \cdot 0.0 + 0,0 \cdot 0.0 \quad = \quad 0.7 \end{aligned}$$

In der Regel besitzen die *Patch Vektoren* mehrere Eigenschaften, ein realistisches Beispiel für die Ähnlichkeit von zwei *Patch Vektoren* berechnet sich nach:

$$\begin{aligned} \vec{v_1} &= (0.31, 0.00, 0.49, 0.61) \\ \vec{v_2} &= (0.00, 0.45, 0.82, 0.76) \\ S_{AB} &= 0.31 \cdot 0.00 + 0.00 \cdot 0.45 + 0.49 \cdot 0.82 + 0,61 \cdot 0.76 \quad = \quad 0.87 \end{aligned}$$

Die Clique-Ähnlichkeitsanalyse kann jetzt auf zwei Arten durchgeführt werden, je nachdem welche Information man zum Aufbau des Clique Produktgraphen bzw. während des Bewertungsschrittes verwendet.

1. Während des Aufbaus des Clique-Eingabegraphen werden nur solche Pseudozentren als ähnlich betrachtet, die einen vergleichbaren Pseudozentrumstyp aufweisen. Dieser Schritt entspricht dem ursprünglichen Ansatz. Während des Scorings wird aber zur Bewertung von ähnlichen Oberflächenbereichen, Information über vergleichbare Oberflächen verwendet. Nach der Überlagerung von zwei Bindetaschen anhand der gefundenen Cliquelösung werden nur solche Pseudozentrenpaare im Scoring berücksichtigt, bei denen das Skalarprodukt der beiden *Patch Vektoren* größer als 0.7 ist. Diese Parametereinstellung wird im Folgenden als ScorePatch (SPatch) bezeichnet.

2. Während des Aufbaus des Clique-Eingabegraphen werden nur solche Pseudozentren als ähnlich betrachtet, die vergleichbare Oberflächenbereiche aufweisen.

Auch für das Scoring bei der Bestimmung der gemeinsamen Oberfläche werden nur solche Pseudozentrenpaare berücksichtigt, deren *Patch Vektoren* als ähnlich angesehen werden. Diese Parametereinstellung wird im Folgenden als CliqueScorePatch (CSPatch) bezeichnet.

Durch die Einführung neuer Pseudozentren (siehe Abschnitt 3.2) und die Verwendung von Oberflächenpunkten mit multiplen Eigenschaften im Vergleich von Bindetaschen muß überprüft werden, wie gut Cavbase nun in der Lage, Ähnlichkeiten zwischen Bindetaschen zu entdecken. Es ist durchaus möglich, daß durch die genannten Erweiterungen, die Ergebnisse der Ähnlichkeitsanalyse verschlechtert werden. Zur Validierung wurden Ähnlichkeitsanalysen mit neun Proteinbindetaschen und eine Clusteranalyse mit drei verschiedenen Datensätzen durchgeführt. Vier verschiedene Parametereinstellungen wurden untersucht: Original, OriginalNeu, SPatch und CSPatch:

1. Der Original Ansatz (Original). Die verwendeten Parameter in der Beschreibung der Bindetaschen und während des Vergleichsprozeßes entsprechen dem ursprünglichen Cavbase Ansatz[Schmitt, 2000]. Die Ergebnisse der drei folgenden Ansätze werden mit diesem Ansatz verglichen.

2. Standard-Ansatz mit erweiteter Pseudozentrendefinition (OrignialNeu). Die Aminosäuren werden basierend auf der erweiterten Pseudozentrendefinition in Pseudozentren übersetzt (siehe Abschnitt 3.2). Die Bindetaschenvergleiche werden nach dem ursprünglichen Cavbase Ansatz durchgeführt.

3. Die nächsten beiden Ansätze verwenden die Oberflächenbereiche mit multiplen Eigenschaften, um festzustellen, ob Pseudozentrenpaare oder Oberflächenbereiche ähnlich sind. Beide Ansätze arbeiten mit der erweiterten Pseudozentrendefinition. Sie unterscheiden sich dadurch, zu welchem Zeitpunkt Information über Oberflächenbereiche mit multiplen Eigenschaften benutzt wird. In der ersten Variante wird der Clique Produktgraph analog zum Original Ansatz durch Pseudozentren gleichen Typs aufgebaut, während im Scoring zwei Oberflächenbereiche dann als ähnlich betrachtet werden, wenn sie über ähnliche *Patch Vektoren* verfügen. Diese Einstellung wird im Folgenden als ScorePatch (SPatch) bezeichnet.

   Im zweiten Fall werden Pseudozentrenpaare dann in den Clique Produktgraphen eingefügt, wenn das Skalarprodukt der zugehörigen *Patch Vektoren* größer als 0.7 ist. Auch im Scoring werden Paare von Pseudozentren dann weiter

berücksichtigt, wenn die zugehörigen Oberflächen als ähnlich angesehen werden (CliqueScorePatch (CSPatch)).

Die neun Bindetaschen in der Vergleichsanalyse wurden gegen einen Datensatz von 3652 Bindetaschen verglichen. Alle Bindetaschen aus dem 3652er Datensatz, die sowohl eine ähnliche Klassifizierung wie die Anfragetasche nach der SCOP-Datenbank erhalten, als auch mit dem Original Ansatz unter den besten 10% der Ähnlichkeitsanalyse gefunden wurden, werden als *echte Hits* bezeichnet. Es wurde nun geprüft, wie viele dieser *echten Hits* mit den drei alternativen Parametereinstellungen (OriginalNeu, SPatch, CSPatch) unter den ersten 10% der besten Treffer des jeweiligen Vergleichs gefunden wurden. In Tabelle 3.1 sind die nach der SCOP-Datenbank verwandten Strukturen und die Anzahl der gefundenen *echten Hits* für die verschiedenen Parameterdefinitionen aufgelistet. Cavbase ist mit allen drei neuen Parametereinstellungen in der Lage, signifikant ähnliche Bindetasche zu detektieren und findet die verwandten Bindetaschen im Datensatz auf den ersten Rängen. Die Ergebnisse stehen in Einklang mit den Resultaten, die mit dem ursprünglichen Ansatz (Original) gewonnen wurden. Sowohl die Verwendung der neuen Pseudozentren Definition als auch der Einsatz der Information über ähnliche Oberflächenbereiche während der Ähnlichkeitsanalyse geben konsistente Ergebnisse. Im Fall der Parametereinstellung CliqueScorePatch werden für besonders große Bindetaschen weniger Treffer unter den ersten 10% des Datensatzes gefunden. Verwendet man die Information über ähnliche Oberflächenbereiche zum Aufbau des Clique-Eingabegraphen, werden sehr viele mögliche Paare von Pseudozentren aufgrund ähnlicher Oberflächenbereiche als gleich angesehen und in den Clique Produktgraphen eingefügt. Daraus folgt, daß der abgetastete Suchraum unter Verwendung von Standardparametern (100 Cliquelösungen) zu klein ist und man im Fall von sehr großen Bindetaschen unter Verwendung dieses dritten Ansatzes weniger Treffer detektiert.

Die Clusteranalyse wurde nach demselben Protokoll, das in Abschnitt 5.2 beschrieben ist, durchgeführt. Als Datengrundlage wurde ein Datensatz von Proteinkinasen, Carboanhydrasen und ein Satz von 113 Bindetaschen funktionell diverser Proteine aus 16 verschiedenen EC-Familien verwendet. Die Clusteranalyse liefert mit allen drei Parametereinstellungen konsistente Ergebnisse (Daten nicht gezeigt).

Die Laufzeiten für die Parametereinstellungen SPatch und CSPatch sind im Vergleich zum Original Ansatz 1.5 bis 3 mal höher. Sie sind in erster Linie davon abhängig, wie groß und dicht verbunden der Clique Produktgraph ist, d.h. je größer die untersuchten Taschen sind. In einer Zusammenarbeit mit Prof. Hüllermeier und Nils Weskamp

Tab. 3.1 Ergebnisse der Validierungstudie unter Verwendung von verschiedenen Parametern während der Ähnlichkeitsanalyse. Angegeben ist das verwendete Protein und die Proteinfamilie sowie die Anzahl der Bindetaschen der entsprechenden SCOP Superfamilie, die unter den ersten 10% der besten Lösungen gefunden wurden. Die Resultate sind für alle vier verwendeten Parametereinstellungen gezeigt (siehe Text).

Proteinfamilie	PDB Code	Hits Original	Hits OriginalNeu	Hits SPatch	Hits CSPatch
Protein Kinase A	1atp	14	12	12	12
Alkohol Dehydrogenase	1b15	179	177	178	167
Carboanhydrase	1cil	62	62	62	62
Chorismat Mutase	1ecm	6	6	6	6
Acetylcholinesterase	1eve	20	19	20	17
HIV-Protease	1hiv	89	89	89	87
Ribonuklease A	1ruv	14	14	14	14
Trypsin	1tpo	204	203	202	203
Thermolysin	3tmn	47	47	47	47

(Universität Marburg, Fachbereich Informatik) konnte eine weitere Beschleunigung des Vergleichsprozeßes durch den Einsatz eines kombinierten Clique-Hashing Verfahrens erreicht werden [Weskamp et al., 2004]. Dieser Ansatz verbindet die Vorteile einer Cliquedetektion mit *Geometric Hashing* Methoden. In einer Indexstruktur wird Information über lokale Substrukturen einer bestimmten Größe (in vorliegendem Fall: Dreiecke aus Pseudozentren) verwaltet, die sehr schnelle Suchen nach lokalen Übereinstimmungen in zwei Bindetaschen erlaubt. Diese lokalen Treffer werden zu größeren Treffern zusammengefasst. Dadurch werden nur dann Knoten in den Clique Produktgraphen eingefügt, die eine ähnliche Umgebung besitzen. Durch diesen Filterschritt wird die Zahl der falsch positiven Treffer deutlich reduziert und der Cliquealgorithmus kann dadurch einen wesentlich schneller den Produktgraphen durchsuchen. Es bietet sich an, dieses Verfahren zur Ähnlichkeitssuche einzusetzen, wobei die physikochemische Modellierung der Bindetaschen auf den multiplen Oberflächenbereichen aufsetzt. Dadurch wird der Nachteil der längeren Laufzeit ausgeglichen.

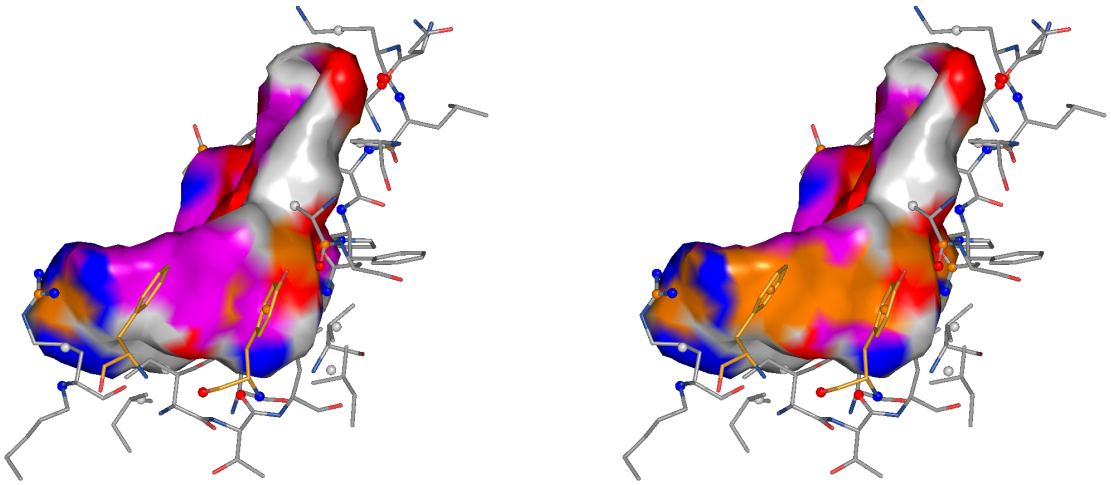
Es ist wünschenswert, daß ein Verfahren zum Vergleich von Proteinbindetaschen unerwartete Ähnlichkeiten zwischen nicht-verwandten Proteinen detektiert. Eine generelle Schwierigkeit im Bewerten von Ähnlichkeiten zwischen nicht-verwandten Bindetaschen ist aber, daß nur wenige validierte Beispiele für diese Art Verwandtschaften bekannt sind. So werden Literatur-bekannte Fälle, wie die Ähnlichkeiten von Bindetaschen der Trypsin- und Subtilisin-Familie oder von Chorismatmutasen verschiedener Spezies, mit

allen Parametereinstellungen detektiert. Es können also bekannte Ähnlichkeiten sehr gut identifiziert werden. Interessant sind aber die Fälle, in denen sich die neuen Parametereinstellungen von dem ursprünglichen Ansatz unterscheiden. Zeigen zwei Proteine keine globale Ähnlichkeit, sondern beschränkt sich ihre Ähnlichkeit auf lokale Bereiche in der Bindetasche, ist es schwierig, die Relevanz dieser Ähnlichkeiten abzuschätzen. In dieser sogenannten *twilight-zone* zeigen sich die Unterschiede zwischen den verwendeten Parametereinstellungen. Prädiktive Aussagen über mögliche Ähnlichkeiten von Bindetaschen nicht verwandter Proteine, wie sie beispielsweise mit der Parametereinstellung CSPatch gefunden wurden, sind nur dann sicher zu belegen, wenn eine experimentelle Überprüfung erfolgt ist.

3.3.2 Repräsentation der Pi-Wechselwirkung von aromatischen Aminosäuren

Im vorherigen Abschnitt wurde die Annotation von Oberflächenpunkten mit mehreren physikochemischen Eigenschaften beschrieben. Es kann aber auch vorkommen, daß Oberflächenpunkte mit keiner Eigenschaft annotiert werden, d.h. sie besitzen kein Pseudozentrum in Interaktionsdistanz. Typischerweise findet man solche Bereiche am Rand der Bindetaschen, wo sich die Oberfläche dem Solvent zuwendet. Es existieren aber auch unerwartete Fälle, in denen die Oberflächenbereiche ohne zugewiesene Eigenschaft in der Bindetasche verbleiben. Eine Ursache ist die unzureichende Übersetzung der Pi-Wechselwirkung von Seitenketten aromatischer Aminosäuren. Während der Überprüfung, ob ein Pseudozentrum seine Eigenschaft auf die Oberfläche ausrichten kann, werden nur solche Pi Pseudozentren berücksichtigt, die einen Winkel zwischen Standard- und Richtungsvektor von kleiner 60^o einschließen. Dabei werden nur *Stacking* Geometrien berücksichtigt, bei denen zu benachbarten Aromaten oder anderen funktionellen Gruppen eine stapelförmige Anordnung (*Stacking*) auftritt. T-förmige Geometrien (*Edge-to-Face*), wie sie häufig zwischen Aromaten mit zahlreichen C-H Gruppen auftreten, werden mit diesen Parametereinstellungen nicht erfasst. Eine Analyse kristallographisch aufgeklärter Protein-Ligand Komplexen hat aber gezeigt, daß man sowohl T-förmige Geometrien (*Edge-to-Face*) als auch $\pi - \pi$ parallel-verschobene Stapel-förmige Geometrien zwischen aromatischen Ringen beobachten kann, ohne das eine besondere Präferenz für eine bestimmte der beiden Geometrien auftritt [Meyer et al., 2003]. Um deshalb auch T-förmige Geometrien zu berücksichtigen, wird zwischen Pi Pseudozentren aus der Seitenkette aromatischer Aminosäuren und Pi Pseudozentren aus der

Peptidbindung unterschieden. Im Falle der Pi Seitenketten Pseudozentren aromatischer Aminosäuren wird der zulässige Winkel zwischen Standard- und Richtungsvektor auf 100^o gesetzt. Dadurch ist nun auch die Erfassung von T-förmigen Geometrien möglich. Der Einfluß des *cut-off* Winkels für Pi Pseudozentren ist in Abbildung 3.6 am Beispiel einer Hydrolase gezeigt. Durch die Verwendung verschiedener *cut-off*-Werte ist eine physikochemisch sinnvollere Beschreibung der Bindetasche möglich.



(I) **(II)**

Abb. 3.6 Berücksichtigung der Seitenketten von aromatischen Aminosäuren in der Annotierung von Proteinbindetaschen. Am Beispiel der Bindetasche einer Hydrolase (PDB Code 1tum) wird der Einfluß verschiedener Parameter bei der Zuweisung der Exposition von Pseudozentren zur Proteinoberfläche gezeigt. In (I) sind zwei Phenylalanine (Kohlenstoffe in orange gefärbt) dargestellt, die eine T-förmige Wechselwirkungsgeometrie in Richtung auf die Bindetaschenoberfläche ausrichten. Bei der ursprünglichen Festsetzung der Winkelabhängigkeiten wird dort keine Interaktion der Oberfläche zugewiesen (violette Bereiche). Im Fall des linken Phenylalanin wird das Pseudozentrum vom Typ Aromatisch herausgefiltert, da es seine Eigenschaft nicht auf die Oberfläche exponieren kann. Standard- und Richtungsvektor schließen einen Winkel ein, der größer als 60 Grad ist. Erlaubt man T-förmige Geometrien, dann werden beide Pseudozentren erkannt und ihre Eigenschaft auf die Bindetaschenoberfläche exponiert (II).

3.4 Validierung der Bindetascheneigenschaften mit wissensbasierten Ansätzen

In diesem Abschnitt wird der Vergleich der Eigenschaftsrepräsentation in Cavbase mit den wissensbasierten Methoden Superstar und Drugscore vorgestellt. Ziel dabei ist es, die bisherige Bindetaschenrepräsentation in Cavbase zu überprüfen und eine möglichst genaue Beschreibung der Bindetaschenoberfläche zu erreichen.

3.4.1 Qualitative Validierung mit Superstar

In einer initialen Studie wurden katalytische Bindetaschen von neun diversen Proteinen[5] untersucht. Superstar [Verdonk et al., 1999, 2001] ist ein Verfahren, um Interaktionsstellen in Proteinbindetaschen zu identifizieren, indem Wahrscheinlichkeitsdichten für das Auftreten von bestimmten Ligandatomen in der Bindetasche als sogenannte *Hotspots* visualisiert werden. Es handelt sich um einen wissensbasierten Ansatz, da nur experimentelle Informationen über nichtbindende Wechselwirkungen aus Kristallstrukturen verwendet werden. Die Datenbank IsoStar liefert somit die Datengrundlage. Die Bindetasche wird in Segmente entsprechend den Isostar-Zentralgruppen zerlegt und die zugehörigen Verteilungen (Scatterplots) aus Isostar auf diese Segmente abgebildet. Schließlich werden die Scatterplots durch Superstar überlagert, in Dichteverteilungen übersetzt und in der Bindetasche visualisiert.

Für die Analyse mit Superstar wurden alle Wassermoleküle und gebundenen Liganden in den Proteineinträgen entfernt und Wasserstoffatome an das Protein addiert. Für die so vorbereiteten Proteine wurde unter Verwendung von fünf Sonden[6] *Hotspot* Felder mit Superstar unter Verwendung von Standardparametern berechnet. Die Gitterweite wurde auf 0.5 Å gesetzt und vergrabene Aminosäuren visuell in der Bindetasche als Startaminosäuren für den Bindetaschendetektionsalgorithmus in Superstar definiert.

[5] Verwendete Proteine: HIV Typ 1 Reverse Transkriptase (3hvt), Ribonuklease T1 (6rnt), Pankreas Elastase (4est), Hummer Enolase (1pdz), Chorismat Mutase (2cht), Glutamin Synthetase (2lgs), 26-10 Fab-Digoxin Komplex (1igj).

[6] Folgende Superstar Probensonden wurden verwendet; in Klammern ist die physikochemische Eigenschaft genannt, die mit der Sonde modelliert werden soll: Amino-Stickstoff (Wasserstoffbrücken-Donor), Alkohol-Sauerstoff (Wasserstoffbrücken-Akzeptor, -Donor), Carbonyl-Sauerstoff (Wasserstoffbrücken-Akzeptor), Aromatischer Kohlenstoff (aromatische Wechselwirkung), Aliphatischer Kohlenstoff (aliphatische Wechselwirkung))

Die so erhaltenen *Hotspot*-Felder wurden mit Pymol graphisch dargestellt, wobei eine Konturierung auf den Niveaus 2.0, 4.0 und 8.0 erfolgte, d.h. die Wahrscheinlichkeit in einem Volumenelement eine bestimmte Gruppe zu finden ist zwei-, vier- oder achtmal höher als die durchschnittliche Wahrscheinlichkeit, sie dort anzutreffen. Bei der graphischen Auswertung wurde darauf geachtet, wie gut die entsprechenden Superstar *Hotspot*-Felder mit den jeweiligen Oberflächenzuweisungen in Cavbase übereinstimmen. Aus den Ergebnissen dieser Untersuchungen können folgende Schlüsse gezogen werden:

- Cavbase detektiert in 100% der Fälle die Bindestelle, in der auch der Ligand bindet. Superstar ist in acht von neun Fällen in der Lage *Hotspot* Felder in der Nähe der Ligandbindestelle zu generieren. Damit ist in diesen Fällen ein Vergleich beider Verfahren möglich.

- Ein visueller Vergleich der Superstar *Hotspots* mit den von Cavbase zugewiesenen Bindetaschenoberflächen eines vergleichbaren physikochemischen Typs zeigt eine gute Übereinstimmung in beiden Darstellungen. Superstar *Hotspot* Felder werden an den entsprechenden Cavbaseoberflächen detektiert und es werden in Cavbase keine Oberflächen generiert, die von physikochemisch abweichenden Superstarfeldern besetzt werden. Unterschiede ergeben sich aber in der genauen Lokalisation und Ausdehnung der Superstar *Hotspot* Felder und den von Cavbase zugewiesenen Oberflächensegmenten.

- Die Auswahl der Startaminosäuren für den Taschendetektionsalgorithmus von Superstar nimmt hat in manchen Fällen einen starken Einfluss auf die errechneten Felder. So werden in einigen Fällen in Bereichen 'hinter' der eigentlichen Bindetaschenoberfläche Felder angezeigt. Um die Cavbaseoberfläche mit den Superstarfeldern abzustimmen, wurden verschiedene Aminosäuren als Startaminosäuren ausgewählt. Superstar konnte nicht in allen Fällen *Hotspots* im Bereich der Bindetasche generieren. Dies erschwert den Vergleich wesentlich.

Aus diesen Gründen wurde nach einem Verfahren gesucht, daß generischer und besser automatisierbar zu verwenden ist wie Superstar. In unserer Arbeitsgruppe bestehen gute Erfahrungen mit dem Programm Drugscore. Gerade die bessere Automatisierbarkeit spricht für die Verwendung von Drugscore.

3.4.2 Qualitative Validierung mit Drugscore

Drugscore ist eine wissensbasierte Bewertungsfunktion, die Geometrien und Affinitäten von Protein-Ligand-Komplexen bewertet[Gohlke et al., 2000a,b; Gohlke and Klebe, 2001]. Zusätzlich ist auch eine Identifizierung von günstigen Wechselwirkungsbereichen in der Bindetasche möglich. Diese können graphisch visualisiert werden (Drugscore *Hotspots*). Für die Validierung der Eigenschaftsrepräsentation in Cavbase ist es wichtig, diejenigen Atomtypen aus den Drugscore Paarpotentialen auszuwählen, die die physikochemischen Eigenschaften der Pseudozentren und Oberflächenbereiche in Cavbase wiedergeben. Die ausgewählten Sondenatomtypen sollten außerdem in den Paarpotentialen ausreichend populiert sein. Gohlke et al. [Gohlke et al., 2000b] verwenden zur Validierung der Drugscore-*Hotspots* die in Tabelle 3.2 aufgeführten Atomtypen als Sondenatome. Diese werden nach ihrem hydrophoben und hydrophilen Charakter sowie nach ihren physikochemischen Interaktionsmöglichkeiten charakterisiert. Gerade letztere lassen einen unmittelbaren Vergleich mit den Cavbase-Eigenschaften zu. Für einige Interaktionseigenschaften existieren mehrere Atomtypen, so kann ein Wasserstoffbrücken-Akzeptor durch die Atomtypen O.3, O.2, O.co2 repräsentiert werden. In dem Vergleich mit Drugscore wurde derjenige Atomtyp ausgewählt, der am besten populiert ist (siehe Tabelle 3.2). In der Validierungsstudie wurden 204 Proteine aus dem CCDC-Astex Datensatz [Nissink et al., 2002] benutzt. Nach folgendem Protokoll wurden die Drugscore-*Hotspots* berechnet:

- Der Ligand wird aus der ursprünglichen PDB-Datei entfernt und im Mol2-Format als Eingabe für Drugscore verwendet.
- Die Bindetasche wird als ein Bereich von 6.0Å um den Liganden definiert.
- Das Gitter, das die Bindetasche in Cavbase umgibt, wird direkt zur Berechnung von *Hotspots* in Drugscore benutzt. Dadurch ist ein direkter Vergleich der Drugscore *Hotspots* mit den Cavbase Bindetaschenoberflächeregionen möglich, da beide Gitter, auf denen die Eigenschaftswerte abgelegt sind, identische Dimensionen besitzen.

Bei einer visuellen Auswertung von *Hotspots* hat die Wahl des geeigneten Konturniveaus einen großen Einfluß auf die Interpretation der Ergebnisse. Nach der Berechnung der Drugscore-*Hotspots* werden diese auf Werte zwischen 0 und 100.0 skaliert, wobei Werte von 100.0 sehr günstige Bereiche für die Wechselwirkung mit dem verwendeten

Tab. 3.2 Verwendete Probenatomtypen in der Drugscore *Hotspot* Analyse zur Validierung der Oberflächeneigenschaften in Cavbase.

Drugscore Atomtyp	Cavbase-Eigenschaft	hydrophob/ hydrophil	physikochemischer Interaktionstyp
C.3	Aliphatisch	hydrophob	aliphatisch
C.ar	Pi	hydrophob	aromatisch
O.3	Donor-Akzeptor	hydrophil	Wasserstoffbrücken Donor/Akzeptor
O.2	Akzeptor	hydrophil	Wasserstoffbrücken Akzeptor
N.3	Donor	hydrophil	Wasserstoffbrücken Donor, protoniert

Sondenatom in der Bindetasche darstellen. Allgemein kann davon ausgegangen werden, daß ein Konturniveau von 90% (d.h. es werden nur solche Gitterpunkte konturiert, die einen Wert größer 90 haben) für eine *Hotspot*-Analyse geeignet ist. Die relative Höhe der einzelnen Werte ist aber von dem betrachteten Protein aus gesehen unterschiedlich und hängt ebenso von den verwendeten Atomtypen ab. Deshalb wurde zur Bestimmung eines geeigneten Konturniveaus eine iterative Vorgehensweise verwendet. Für ein Gitter wurde das Konturniveau so lange verändert, bis 0.6% der Gitterpunkte für den jeweiligen Atomtyp konturiert wurden. Dieses Vorgehen wurde empirisch validiert und liefert für eine *Hotspot*-Analyse zu große Konturflächen, kann aber zur Validierung von Bindetascheneigenschaften - und damit zur Klärung der Frage, was für physikochemische Eigenschaften überhaupt in bestimmten Bereichen der Proteinbindetaschen anzutreffen sind - verwendet werden.

Die graphische Auswertung und der Vergleich der Drugscore *Hotspots* mit den Cavbaseoberflächenregionen liefert eine gute Übereinstimmung zwischen beiden Beschreibungen. Drugscore berechnet *Hotspots* für bestimmte Atomtypen, die in der Bindetasche lokalisiert sind. Die von Cavbase generierte Bindetaschenoberfläche ist ähnlich einer Connolly Oberfläche 'näher' an den Aminosäuren lokalisiert. Projiziert man die Eigenschaften der Drugscore *Hotspots* auf die Bindetaschenoberfläche, erhält man eine gute Übereinstimmung zwischen beiden Beschreibungen. Gerade die Hotspots für

gerichtete Wechselwirkungen (Wasserstoffbrücken-Wechselwirkungen) werden mit den entsprechenden Oberflächenregionen aus Cavbase gut überlagert (siehe Abbildung 3.7).

Drugscore findet in bestimmten Volumenbereichen der Bindetasche nicht nur *Hotspots* für einen einzigen Atomtyp, sondern *Hotspots* mehrerer Atomtypen gleichzeitig (Konturierung der 0.6% Gitterpunkte). In Abschnitt 3.3 wurden Oberflächenpunkte mit multiplen Eigenschaften vorgestellt. Eine Beobachtung war, daß Oberflächenpunkte mehrere physikochemische Eigenschaften besitzen, da sie sich in Interaktionsdistanz zu mehreren Pseudozentren befinden. Auch die Drugscore *Hotspot* Analyse unterstützt diese Annahme.

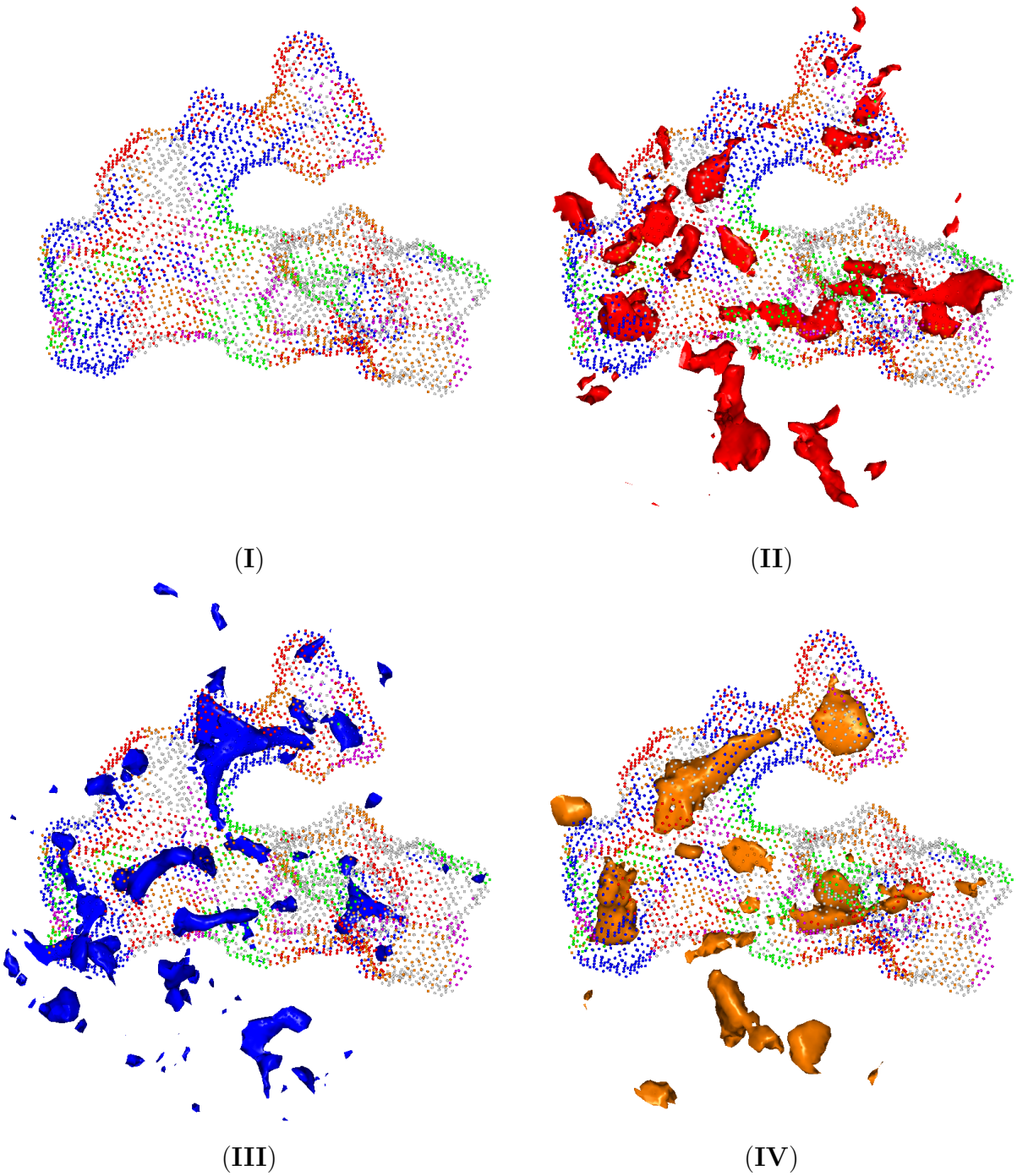


Abb. 3.7 Vergleich von Drugscore *Hotspots* mit der Cavbase Bindetaschenoberfläche am Beispiel einer Dihydro-Orotat-Dehydrogenase (PDB Code 1d3h). In (I) bis (IV) ist die Bindetasche einer Dihydro-Orotat-Dehydrogenase gezeigt. (I) stellt **nur** die Cavbase Oberfläche als Punkte dar. Die Oberfläche ist entsprechend der fünf physikochemischen Eigenschaften eingefärbt, die in Cavbase verwendet werden (siehe Abbildung 5.1). In (II) bis (IV) sind zusätzlich zur Cavbase Oberfläche noch drei Drugscore *Hotspot*-Felder dargestellt. Die Farbcodierung der Drugscore *Hotspot* Felder ist so gewählt, daß sie die gleiche Farbe wie die Cavbase Bindetaschenoberfläche der entsprechenden physikochemischen Eigenschaft besitzen: Drugscore Atomtyp N.3 (rot, II), O.2 (blau, III) und C.ar (orange, IV). Drugscore *Hotspots*, die gerichtete Wechselwirkungen wiedergeben, werden sehr gut mit den entsprechenden Cavbase Oberflächen überlagert. Das Konturierungsniveau ist so gewählt, daß 0.6% der Gitterpunkte, die die günstigsten Wechselwirkungsbereiche für den jeweiligen Atomtyp darstellen, für den jeweiligen Atomtyp gefärbt sind.

3.5 Optimierungen von Bindetaschenvergleichen

Eine Voraussetzung für die Ähnlichkeitsanalyse großer Datensätzen ist ein Vergleichsverfahren, das Gegenüberstellungen von Bindetaschen effizient und schnell berechnen kann. Mit der Methode Cavbase, die zwei Bindetaschen mit Hilfe eines Clique Algorithmus [Bron and Kerbosch, 1973] vergleicht und die Ähnlichkeit anhand der Überlappung beider Bindetaschenoberflächen bewertet, konnte ein solches Verfahren etabliert werden [Schmitt, 2000].

Ein Cliquealgorithmus findet den größten gemeinsamen Subgraphen (Clique) in zwei Graphen. Eine Clique ist als größter Subgraph eines Graphen definiert, in dem alle Knoten miteinander verbunden sind. Bindetaschen werden in Cavbase als Graph mit den Pseudozentren als Knoten repräsentiert. Das Auffinden gemeinsamer Bereiche in zwei Bindetaschen kann als die Detektion gemeinsamer Subgraphen formuliert werden. Um zwei Bindetaschen **G1** und **G2** mit dem Cliquealgorithmus vergleichen zu können, muß zuerst der Clique-Eingabegraph **P** (Produktgraph) gebildet werden. Ein Knoten (u_1, v_1) in **P** besteht aus zwei Pseudozentren (jeweils einem aus jeder Bindetasche). Es wird immer dann ein Knoten in **P** eingefügt, wenn ein Paar von Pseudozentren aus den beiden Bindetaschen den gleichen Typ besitzt. Nachdem alle möglichen Kombinationen von Pseudozentrenpaare enumeriert worden sind, werden zwei Knoten (u_1, v_1) und (u_2, v_2) in **P** verbunden, wenn die Distanz zwischen den entsprechenden Pseudozentrenpaaren in den beiden Bindetaschen ähnlich ist, d.h. die Differenz der beiden Distanzen innerhalb bestimmter Toleranzgrenzen liegt. Um sich im Vergleich der Bindetaschen auf lokale Muster zu konzentrieren, werden nur solche Paare von Pseudozentren berücksichtigt, die nicht länger als 12.0Å (d_{max}) entfernt sind. Die Toleranzgrenze für die Abweichung der Distanzen wurde auf 2.0Å (d_{diff}) gesetzt. Gleichung 3.5 fasst noch einmal die Bedingungen zusammen, die erfüllt sein müssen, damit zwei Knoten (u_1, v_1) und (u_2, v_2) in **P** verbunden werden können.

$$P((u_1, v_1), (u_2, v_2)) = \begin{cases} 1 \quad \text{wenn} & \begin{array}{l} \mid d(u_1, v_1) - d(u_2, v_2) \mid \leq d_{diff} \quad \wedge \\ d(u_1, v_1) \leq d_{max} \quad \wedge \\ d(u_2, v_2) \leq d_{max} \end{array} \\ 0 \quad \text{sonst.} & \end{cases} \tag{3.5}$$

Eine Strategie zur Beschleunigung von Bindetaschenvergleichen besteht darin, die Anzahl der verbundenen Knoten in **P** zu reduzieren. Große Produktgraphen, deren Knoten stark miteinander verbunden sind, führen zu sehr hohen Laufzeiten. Deshalb sollen nur solche Knoten in **P** miteinander verbunden werden, deren Pseudozentrenpaare neben einer ähnlichen Distanz und dem gleichen Pseudozentrumstyp auch noch andere Gemeinsamkeiten zeigen. Dadurch soll die Wahrscheinlichkeit erhöht werden, daß es sich bei der Übereinstimmung nicht nur um eine zufällige Ähnlichkeit handelt, sondern die zugrundeliegenden Pseudozentrenpaare Teil eines größeren Musters sind. Verschiedene Filterschritte sind zur Reduzierung möglicher Pseudozentrenpaare durchgeführt worden [Schmitt, 2000]. So wurden zwei Knoten in **P** nur dann miteinander verbunden, wenn sie eine vergleichbare räumliche Orientierung besitzen. Diese wurde durch den Winkel zwischen Standard- und Richtungsvektor der jeweiligen Pseudozentren ausgedrückt. Liegt die Differenz der Winkel über einem bestimmten *cut-off*-Wert, werden beide Pseudozentren nicht als ähnlich angesehen. Filterschritte, die alleine auf geometrischen Gesichtspunkten beruhen, haben aber den Nachteil, daß sie nicht sicher richtige von falsch positiven Übereinstimmungen von zwei Pseudozentrenpaaren erkennen können. Eine Gefahr besteht zudem darin, daß die Information über sinnvolle Überlagerungen an einem zu frühen Punkt nicht weiter verwendet wird. Deshalb wurden die Parametereinstellungen für die Clique-Detektion *weich* gesetzt und als ein entscheidendes Bewertungskriterium der Überlappungsgrad von ähnlichen Oberflächen angesehen. Der Scoringschritt hat die Aufgabe, zwischen sinnvollen und nicht-sinnvollen Überlagerungen zu unterscheiden. Eine Optimierung, die die Größe des Produktgraphen **P** verringert ohne dabei auf richtige Lösungen zu verzichten, ist die Anpassung der Toleranzgrenze d_{diff} an die entsprechende Länge zwischen den betrachteten beiden Pseudozentrenpaaren. Im ursprünglichen Ansatz beträgt die Toleranzgrenze 2.0Å. Eine solche Toleranzgrenze d_{diff} ist bei langen Distanzen sinnvoll. Bei Pseudozentrenpaaren, die nur durch kurze Distanzen getrennt sind, ist eine solche Toleranzgrenze jedoch zu groß bemessen. Deshalb wurde die erlaubte Differenz Distanz-abhängig angepasst. In Tabelle 3.3 sind die in dieser Arbeit verwendeten Parameter für d_{diff} gezeigt.

Der Clique Algorithmus findet basierend auf dem Produktgraphen **P** eine vorgegebene Zahl (100) bestmöglicher Cliquen, die alle eine mögliche Überlagerung der beiden Bindetaschen darstellen. Von den zur Übereinstimmung gebrachten Pseudozentren der Cliquelösung ausgehend, kann eine Transformation (Translation und Rotation) berechnet werden, anhand der die gesamten Bindetaschen überlagert werden. Um zu bewerten, welche von diesen Cliquelösungen beide Bindetaschen nach Form und physikochemi-

Tab. 3.3 Distanzabhängige Anpassung von d_{diff} an die Distanz zwischen zwei Pseudozentren, die während des Aufbaus des Clique-Eingabegraphen verwendet wird.

Distanz zwischen zwei Pseudozentren [Å]	Toleranzgrenze d_{diff} [Å]
> 10.0	2.0
8.0 - 10.0	1.6
4.0 - 8.0	1.2
2.0 - 4.0	0.8
< 2.0	0.6

schen Eigenschaften am besten überlagert, wird nun überprüft, wie gut Oberflächenbereiche gleichen Typs überlappen. Dadurch wird sichergestellt, daß nur solche Cliquen positiv bewertet werden, die neben einer ähnlichen Form (definiert über die initial gefundene Clique) auch noch eine physikochemische Ähnlichkeit zeigen (definiert über überlappende Oberflächenbereiche gleichen Typs).

Im Bewertungsschritt wird eine Schleife über alle Pseudozentrenpaare gleichen Typs durchlaufen (also nicht nur über die Paare, die Teil der Cliquelösung sind) und die Überlappung der Oberflächenpunkte bestimmt. Zwei Oberflächenpunkte zweier Oberflächenbereiche gelten als überlappend, wenn die Distanz zwischen ihnen kleiner als 1.0Å ist. Um schwach überlappende Oberflächenbereiche nicht zu berücksichtigen, werden nur solche Oberflächenbereiche im Scoring verwendet, deren Oberflächen zu mehr als 70% überlappen. Wenn p_{a_i} und p_{b_k} jeweils Punkte eines zugehörigen Oberflächenbereichs sind ($p_{a_i} \in \mathbb{P}_{a_i}$; $p_{b_k} \in \mathbb{P}_{b_k}$), dann berechnet sich R_1 nach [Schmitt, 2000]:

$$R_1 = \sum_v \sigma_v \qquad \text{(für alle } \sigma_v \geq 0.7\text{)} \tag{3.6}$$

$$\begin{aligned} \sigma_v &= \frac{\rho_{a_i} + \rho_{b_k}}{\mid \mathbb{P}_{a_i} \mid + \mid \mathbb{P}_{b_k} \mid} \\ \rho_{a_i} &= \mid \{p_{a_i} \mid d(p_{a_i}; p_{b_k}) \leq 1.0\} \mid \\ \rho_{b_k} &= \mid \{p_{b_k} \mid d(p_{b_k}; p_{a_i}) \leq 1.0\} \mid \end{aligned} \tag{3.7}$$

R_1 (Gleichung 3.6) korreliert mit der Anzahl der gemeinsamen Oberflächenbereiche und mit der Güte, wie diese Oberflächenbereiche überlappen. Die Bewertungsfunktion R_2 (Gleichung 5.1) berücksichtigt neben der absoluten Größe der überlappenden Oberflächen noch zusätzlich die RMS-Abweichung ($rmsd$) der Pseudozentren der ent-

sprechenden Cliquelösung. Dabei werden solche Cliquelösungen schlechter bewertet, bei denen die Pseudozentren eine große Abweichung zeigen.

Zusätzlich zu R_1 und R_2 wurde eine weitere Bewertungsfunktion R_3 (Gleichung (5.1)) entwickelt, die ähnlich dem Tanimoto Index [Godden et al., 2000] berechnet wird. Sie setzt die Anzahl der als ähnlich gefundenen Pseudozentren (n_{ctr}) zu der Größe der jeweiligen Bindetasche ($n_{maxcav1}$ und $n_{maxcav2}$) in Beziehung. R_3 liefert Werte zwischen Null und Eins.

$$\begin{aligned} R_2 &= \frac{R_1 - 0.7\ n}{rmsd} \\ R_3 &= \frac{n_{ctr}}{n_{maxcav1} + n_{maxcav2} - n_{ctr}} \end{aligned} \tag{3.8}$$

Eine Analyse des ursprünglichen Vergleichsverfahrens hat gezeigt, daß besonders zwei Aspekte kritisch für den Zeitbedarf eines Bindetaschenvergleichs sind: zum einem die Größe des Produktgraphen **P** während der Cliquedetektion und zum anderen der Scoringschritt. In dem ursprünglichen Vergleichsverfahren wurde ungefähr 80% der Rechenzeit für das Scoring benötigt.

Anhand der gefundenen Cliquelösung werden die Bindetaschen überlagert und im Bewertungsschritt wird der Grad der Überlappung der Oberflächenbereiche als Ähnlichkeitsmaß bestimmt. Dazu müssen beiden Bindetaschen zuerst basierend auf der gefundenen Transformationsmatrix der Cliquelösung überlagert werden. Eine Bindetasche wird in Cavbase als C++-Objekt (Cavity) repräsentiert. Ein Bindetaschenobjekt enthält wiederum Listen von Pseudozentren und Oberflächenpunkten, die ebenfalls als Objekte in der Programmiersprache C++ realisiert sind. Bevor die Bindetasche anhand der gefundenen Cliquelösung überlagert wird, muss zuerst eine Kopie des Cavity Objektes erzeugt werden. Dabei werden alle Attribute einer Bindetasche, die selbst als Objekte implementiert sind, ebenfalls durch Anlegen einer Kopie neu erzeugt. Eine Profiling-Analyse hat gezeigt, daß gerade aufwendige Rechenoperationen, wie das Kopieren von Objekten (Pseudozentren, Oberflächenpunkte) zeitkritisch sind. Durch Implementieren von speziellen *Copy-Konstruktoren*, die eine schlankere Repräsentation der Cavity Objekte anlegen, werden nur solche Attribute kopiert, die für die Ähnlichkeitsanalyse wichtig sind. Dadurch konnte der Prozeß wesentlich beschleunigt werden.

Der Schritt, der während eines Bindetaschenvergleichs die meiste Zeit benötigte, ist das Scoring. Beide Bindetaschen werden überlagert und es wird überprüft, wie gut die Oberflächenbereiche *aller* erlaubten Kombinationen von Pseudozentren-Paaren gleichen Typs überlappen. In diesem Schritt werden viele unnötige Distanzberechnungen

vorgenommen, was wesentlich zum hohen Zeitaufwand beiträgt. Ein Analyse von Pseudozentrenpaaren, die Teil einer guten Cliquelösung sind und solchen, die während des Bewertungsschritts herausgefiltert werden, hat gezeigt, daß die Oberflächenbereiche von Pseudozentren, die mehr als 4Å entfernt liegen, in der Regel nicht sinnvoll überlagert werden können. Deshalb wurden nur solche Pseudozentrenpaare im Oberflächenscoring berücksichtigt, die weniger als 4Å entfernt liegen. Dabei wird eine große Anzahl an unnötigen Berechnungen eingespart. Weitere potentielle Optimierungen, wie die Berücksichtigung der relativen Orientierung der Pseudozentren (definiert über den Winkel zwischen dem Standard- und Richtungsvektor) während des Scorings waren in ihrer Berechnung selbst zu aufwendig und konnten nicht zur Beschleunigung beitragen.

Durch diese Optimierungen konnte der Vergleich von Bindetaschen um den Faktor 40 beschleunigt werden und es ist nun möglich, Ähnlichkeitsanalysen in sehr großen Datenmengen (Paarvergleiche im Millionenmaßstab) durchzuführen.

3.6 Vergleiche von Bindetaschen mit Hilfe von Bitstrings

Fingerprints (Bitstrings) repräsentieren die strukturelle Information und Gestalt von 3D-Objekten in einem eindimensionalen Textstring und werden in vielen Bereichen der Objekterkennung eingesetzt. Für die Ableitung von Bitstrings aus der Struktur eines chemischen Moleküls existieren eine Vielzahl an Verfahren. Im Allgemeinen wird bei Vorhandensein eines bestimmten Fragmentes im Molekül ein Bit auf dem Bitstring gesetzt. Andere Ansätze kommen ohne vorher definierte Fragmentbibliotheken aus, indem alle prinzipiell denkbaren Fragmente (beispielsweise einer bestimmten Größe) benutzt werden, um den Bitstring zu erzeugen. Mit der Repräsentation durch Bitstrings lassen sich sehr effizient Vergleiche und Ähnlichkeitssuchen von Molekülen durchführen. Bitstrings werden deshalb auf vielen Gebieten des Wirkstoffdesigns eingesetzt und finden Anwendung in der Substruktursuche und Ähnlichkeitssuchen in 3D-Datenbanken von kleinen Molekülen oder zur Identifizierung von Molekülen, die ein bestimmtes Pharmakophor Muster erfüllen (siehe auch [Downs and Willett, 1996], [Leach, 2001] Kapitel 12 und [Gasteiger and Engel, 2003] Kapitel 2.7 und 6.4).

Prinzipiell ist eine Anwendung von Bitstringvergleichen auch auf Bindetaschen denkbar. In einer vorherigen Arbeit [Schmitt, 2000] wurde versucht, zwischen ähnlichen und unähnlichen Bindetaschen mit Hilfe eines neuronalen Netzes zu unterscheiden. Als Eingabevektor für das neuronale Netz dienten Verteilungen der Distanzen zwischen zwei Pseudozentren bestimmter Pseudozentrumstypen. Mit diesem Ansatz war es aber nicht möglich, Bindetaschen verschiedener Proteinfamilien zu separieren. Das neuronale Netz trennte die Bindetaschen eher nach verschiedener Größe als nach Aufbau und physikochemischer Zusammensetzung. Da selbst Bindetaschen eines bestimmten Proteins aufgrund von Proteinflexibilität - je nach zugrundeliegender Kristallstruktur - sich in der Ausdehnung unterscheiden, wurden Bindetaschen einer Proteinfamilie nicht in dem gleichen Neuron gefunden.

Um die Eigenschaften und den strukturellen Aufbau der Bindetasche besser beschreiben zu können, werden in dieser Arbeit deshalb Tripletts von Pseudozentren gebildet und zur Kodierung der Information über eine Bindetasche in einem Bitstring verwendet. Ähnliche Ansätze wurden zur Identifizierung von 3D-Pharmakophormustern in Datenbanksuchen gewählt [Good and Kuntz, 1995; Good et al., 1995]. Prinzipiell ist es natürlich auch möglich, größere Strukturen wie Tetraeder oder noch größere Objekte

zur Erzeugung der Bitstrings zu benutzen (vgl. Dissertation K. Kupas, Universität Marburg). Dies hat den Vorteil, daß mehr Information über lokale Substrukturen kodiert und dadurch die Wahrscheinlichkeit für falsch-positive Treffer minimiert wird. Andererseits wird die Komplexität zur Berechnung und Verwaltung der Bitstrings stark erhöht. Als Kompromiß aus beiden Aspekten wurden Substrukturen der Größe drei benutzt.

Bitstrings für Bindetaschen werden nach folgendem Algorithmus erzeugt:

- Ein Bitstring für Bindetaschen umfasst alle möglichen Dreiecke bestehend aus drei Pseudozentren, deren Seiten eine gewisse Länge besitzen. Dazu werden alle möglichen Kombinationen aus drei Pseudozentren gebildet. Aus den fünf möglichen Pseudozentrumstypen kann man 35 Kombinationen von 3er-Tupeln (z.B. Akzeptor,Aliphatisch,Donor-Akzeptor) bilden. Dies entspricht im Urnenmodell dem Ziehen von k Kugeln aus n Möglichen, wobei man Wiederholungen zulässt (Tripel dürfen mehrere Pseudozentren gleichen Typs enthalten) und die Reihenfolge, in der die Kugeln gezogen werden, nicht beachtet. Die Anzahl der möglichen Kombinationen berechnet sich nach Formel 3.9 :

$$\binom{n}{k} = \frac{(n+k-1)!}{(n-1)! \cdot k!} \tag{3.9}$$

- Für jede dieser 35 Kombinationen wird ein Bitstring einer bestimmten Länge kodiert. Ein kompletter Bitstring für eine Bindetasche besteht aus 35 solcher Bitstrings.

- In der Beschreibung der Bindetasche werden nicht alle möglichen Dreiecke betrachtet. Es werden nur solche Dreiecke berücksichtigt, in denen keine Seitenlänge kleiner als 3 Å und größer als 10 Å ist. Damit sollen nur solche Dreiecke zur Beschreibung der Bindetasche beitragen, die charakteristische lokale Muster der Pseudozentren beschreiben. Ist die Distanz zwischen zwei Pseudozentren kleiner als 3 Å, werden vor allem Abstände von Pseudozentren einer Aminosäure beschrieben. Um aber gerade diskriminierende Muster in einer Bindetasche zu erfassen, wird auf diese Distanzen verzichtet. Die Obergrenze von 10 Å soll die Fokussierung auf lokale Muster in der Tasche garantieren.

- Die Länge der verwendeten Bitstrings richtet sich nach der kodierten Information. Im ersten Fall wird nur der Flächeninhalt (*area*) eines Dreiecks zur Indexierung des Bitstrings verwendet. Der Flächeninhalt eines Dreiecks mit bekannten Sei-

tenlängen (a, b, c) berechnet sich nach Formel 3.10.

$$area(a,b,c) = \frac{\sqrt{(b+c)^2 - a^2} * \sqrt{a^2 - (b-c)^2}}{4} \quad (3.10)$$

Die maximal mögliche Fläche, die ein Dreieck bei gegebenem Umfang aufspannen kann, berechnet sich nach Formel 3.11 [Good and Kuntz, 1995].

$$maxarea(a,b,c) = \sqrt{\frac{(a+b+c)^4}{432}} \quad (3.11)$$

Der Flächeninhalt wird in Bins der Länge 0.5 $Å^2$ kodiert.

- Ein Nachteil der Indexierung die alleine über den Flächeninhalt vorgenommen wird ist, daß Dreiecke unterschiedlicher Gestalt, die denselben Flächeninhalt besitzen, dasselbe Bit auf dem Bitstring setzen. Um zwischen Dreiecken unterschiedlicher Gestalt zu unterscheiden, wird im zweiten Fall zusätzlich zur Fläche des Dreiecks ein Wert für die Gestalt des Dreieckes kodiert. Dieser berechnet sich als Summe der quadrierten Abweichungen der Seitenlängen von den Seitenlängen eines gleichseitigen Dreiecks gleichen Umfangs. Bei gegebenem Umfang p besitzt ein gleichseitiges Dreieck den größtmöglichen Flächeninhalt. Das Maß für die Gestalt eines Dreieckes berechnet sich nach Gleichung 3.12.

$$\begin{aligned} p &= \frac{(a+b+c)}{3} \\ ratio(a,b,c) &= (\sqrt{(p-a)^2} + \sqrt{(p-b)^2} + \sqrt{(p-c)^2}) + 1 \end{aligned} \quad (3.12)$$

- Nach der Erzeugung der Bitstrings werden noch optional die Bits, die in allen Bitstring des betrachteten Datensatzes öfter als ein bestimmter Wert (z.B. 0.7) gesetzt sind, auf Null gesetzt. Damit soll überprüft werden, ob es möglich ist, die diskriminierende Fähigkeit durch diese Gewichtung der einzelnen Bits noch zu verstärken.

Die Ähnlichkeit von Bindetaschenbitstrings für zwei Bindetaschen A und B wird mit dem Tanimoto-Index berechnet 3.13. a ist die Anzahl der gesetzten Bits in Bindetasche A und b die Anzahl der gesetzten Bits in Bindetasche B; c gibt die Anzahl der Bits an, die in beiden Bindetaschen gesetzt sind.

$$S_{AB} = \frac{c}{a+b-c} \quad (3.13)$$

Vergleiche mit Bindetaschenbitstrings lassen sich sehr schnell durchführen. Die Generierung eines Bindetaschenbitstrings dauert weniger als eine Sekunde (Bindetaschen mit

90 Pseudozentren) und muß auch nur einmal für die gesamte Datenbank durchgeführt werden. Ein Vergleich einer Bindetasche gegen einen Datensatz von mehreren Tausend Bindetaschen kann in wenigen Minuten durchgeführt werden. Im Gegensatz dazu benötigt ein Vergleich mit dem Standard Clique Ansatz gegen denselben Datensatz 2 bis 3 Stunden, je nach Größe der untersuchten Bindetasche. Aus diesem Grund sind Vergleiche mit Bindetaschenbitstrings zum schnellen Vorfiltern von großen Datensätzen sehr gut geeignet.

Zur Evaluierung der Bindetaschenbitstrings wurden 16 Bindetaschen gegen einen Datensatz von 6320 Bindetaschen verglichen (siehe Tabelle 3.4). Unter diesen Datensatz wurden Bindetaschen der gleichen SCOP Superfamilie gemischt. Proteine derselben SCOP-Superfamilie zeigen funktionelle Verwandtschaft in den Bindetaschen. So bilden die Proteinkinasen, Serinproteasen der Trypsin-Familie oder die katalytische Domäne von Metalloproteasen mit Zink im aktiven Zentrum Beispiele für SCOP Superfamilien. Alle Bindetaschen wurden ebenfalls mit dem Clique-Algorithmus unter Verwendung von Standard-Parametern verglichen. Es wurden nur die Bindetaschen als *echte Hits* definiert, die von dem Cliquealgorithmus auf den ersten 10% der Ränge gefunden wurden und zur gleichen SCOP-Superfamilie wie die Anfragetasche gehören. Ein Protein derselben SCOP-Superfamilie kann mehrere Bindetaschen aufweisen, die dann alle dieselbe SCOP-Annotation besitzen. Das Protein enthält beispielsweise nur eine Bindetasche mit katalytischen Resten, die für eine Ähnlichkeitssuche relevant ist.

Es wurde nun überprüft, wie gut man die *echten Hits* mit den Bitstringvergleichen im Datensatz wiederfinden kann. Dazu wurden alle Bindetaschen mit den Bitstringvergleichen untersucht und nach der erhaltenen Ähnlichkeit geordnet. Die Ergebnisse der Ähnlichkeitsanalyse sind in Tabelle 3.4 aufgelistet. Gezeigt ist die jeweilige Bindetasche, die Proteinfamilie, die theoretisch auffindbaren Bindetaschen derselben SCOP Superfamilie und der Prozentsatz der *echten Hits*, die unter den ersten 100, 1000, 3200, 5000 Rängen gefunden wurden.

Tab. 3.4 Ergebnisse der Ähnlichkeitsanalyse mit dem Bindetaschenvergleichsalgorithmus unter Verwendung von Bitstring-Vergleichen. Aufgelistet ist die Cavbase ID, die Anzahl der Pseudozentren der Bindetasche, die Proteinfamilie und die EC-Nummer, die Anzahl an echten SCOP Hits und der Prozentsatz an gefundenen *echten Hits* (normiert auf Werte zwischen Null und Eins) unter den ersten 100, 1000, 3200 und 5000 besten Plätzen (Bitstringvergleiche sortiert nach S_{AB})

Cavbase ID	n_{Pseudo}	**Proteinfamilie**	**EC-Nummer**	**Scop Hits.**	**Hits 100**	**Hits 1000**	**Hits 3200**	**Hits 5000**
1axe.4	114	Alkohol Dehydrogenase	1.1.1.1	31	0.65	0.97	1.00	1.00
2acu.1	85	Aldose Reduktase	1.1.1.21	4	0.00	0.75	0.75	0.75
1emd.1	69	Malat Dehydrogenase	1.1.1.37	163	0.10	0.54	0.98	0.99
1gro.4	101	Isozitrat Dehydrogenase	1.1.1.42	10	0.40	0.60	0.90	0.90
3dhe.1	148	Östrogen Rezeptor	1.1.1.62	137	0.01	0.43	0.99	0.99
1bf3.2	161	Hydroxylase	1.14.13.2	79	0.42	0.96	0.99	1.00
1a47.4	57	Glycosyltransferase	2.4.1.19	26	0.42	0.69	0.85	0.96
1atp.2	98	Proteinkinase A	2.7.1.37	28	0.07	0.46	1.00	1.00
2tmk.4	67	Thymidylat Kinase	2.7.4.9	41	0.00	0.44	0.93	0.98
1bs4.2	46	Peptid Deformylase	2.7.4.6	4	0.25	0.75	1.00	1.00
1tpo.1	78	Trypsin	3.4.21.4	34	0.62	0.74	0.91	1.00
1btz.1	73	Trypsin	3.4.21.4	27	0.63	0.78	1.00	1.00
3prk.1	54	Proteinase K	3.4.21.64	3	0.00	0.33	0.67	1.00
6apr.2	69	Rhizopuspepsin	3.4.23. 6	24	0.29	0.50	0.88	1.00
1cil.1	43	Carboanhydrase	4.2.1.1	14	0.93	1.00	1.00	1.00
1xic.1	151	Xylose Isomerase	5.3.1.5	25	1.00	1.00	1.00	1.00

Eine Analyse der Ergebnisse zeigt, daß Vergleiche mit Bindetaschenbitstrings generell sehr gut in der Lage sind, ähnliche Bindetaschen verwandter Proteine zu identifizieren. In allen Fällen wurden fast alle verwandten Bindetaschen gefunden, wenn man die 50% der am besten bewerteten Lösungen (3200 Hits) untersucht. Vergleicht man alle Bindetaschen eines Datensatzes mit Hilfe von Bitstrings und sortiert die Ergebnisse nach der gefundenen Ähnlichkeit, ist es dann ausreichend, die ersten 50% der Ränge der gesamten Vergleiche mit dem aufwendigeren Clique Algorithmus zu vergleichen, um den Großteil (zwischen 90% und 100%) der funktional verwandten Strukturen aufzufinden. Vergleicht man die ersten 5000 besten Platzierungen der Bitstringvergleiche mit dem Cliquealgorithmus (eine Einsparung um 21% der Bindetaschenvergleiche), ist man in der Lage, für 10 der 16 Vergleiche alle *bekannten Hits* und für weitere fünf Fälle über 95% der *echten Hits* zu detektieren.

Ein Vergleich der beiden verwendeten Verfahren zur Erzeugung der Bindetaschenbitstrings - zum einen alleine über den Flächeninhalt sowie zum anderen über den Flächeninhalt und die Gestalt des Dreiecks - zeigt sehr ähnliche Ergebnisse (siehe Tabelle 3.5). Auch die Nicht-Berücksichtigung von Bits, die zu einem bestimmten Prozentsatz in allen betrachteten Bitstring gesetzt sind (in Tabelle 3.5 für größer 70% gezeigt), hat nicht zur Verbesserung der Ergebnisse beigetragen. Dabei wurden verschiedene Prozentwerte ausprobiert, trotzdem konnte keine signifikante Verbesserung in der Ähnlichkeitsanalyse erzielt werden.

Es ist interessant festzustellen, daß die Vergleiche mit den Bindetaschenbitstring robust gegenüber der Größe der verwendeten Anfrage-Bindetasche sind. Die Größe einer Tasche hat einen starken Einfluß auf den errechneten Bitstring, indem die Wahrscheinlichkeit für das Vorhandensein von Tripletts aus Pseudozentren einer bestimmten Größe ansteigt. Es wäre zu erwarten, daß sehr große Bindetaschen generell eine hohe Ähnlichkeit zu der Anfrage-Bindetaschen zeigen. Analysiert man aber die Anzahl der wiedergefundenen *echten Hits* in Abhängigkeit der Bindetaschengröße, so werden sowohl für kleine als auch für sehr große Bindetaschen (Bindetasche 1bf3.2, 1gro.4) sehr gute Ergebnisse erzielt.

Bindetaschenbitstringverfahren stellen deshalb eine sehr effiziente Methode dar, um bei Vergleichen mit großen Datenmengen, solche Bindetaschenvergleiche zu identifizieren, in denen beide Bindetasche keine große Ähnlichkeit zueinander zeigen. Mit den aufwendigeren Cliquevergleichen müssen dann weniger Vergleiche durchgeführt werden. Bitstrings tragen so maßgeblich zur Beschleunigung dieser Vergleiche bei. Vergleicht

Tab. 3.5 Ergebnisse der Ähnlichkeitsanalyse mit Bindetaschenbitstrings unter Verwendung verschiedener Parameter.

Bitstringverfahren	Gewichtung	Prozentsatz der gefundenen *echte Hits*			
		100	1000	3200	5000
Flächeninhalt	nein	0.36	0.68	0.93	0.97
Flächeninhalt	ja	0.37	0.66	0.93	0.97
Flächeninhalt + Gestalt	nein	0.37	0.67	0.92	0.98
Flächeninhalt + Gestalt	ja	0.35	0.66	0.92	0.98

man beispielsweise einen Datensatz von 2000 Bindetaschen gegen sich selbst und betrachtet mit dem Clique Verfahren nur die besten 80% der Bitstringvergleiche (mittlere Rechenzeit 2 sec pro Bindetaschenvergleich (1CPU)), dann verkürzt sich der Zeitverbrauch um ca. 400 h auf 712 h.

4 Analysen zur Bestimmung der Funktion von Proteinen. Vorhersagen von Kreuzreaktivitäten

4.1 Beispiele für die Funktionsanalyse von Proteinbindetaschen

Betrachtungen im Hinblick auf die Funktion eines Proteins können auf verschiedenen Ebenen ansetzen. Man unterscheidet die biologische Funktion von der biochemischen Funktion eines Proteins [Orengo et al., 1999; Thornton et al., 2000; Bartlett et al., 2003]. Die biologische Funktion eines Proteins beschreibt, welche Aufgaben und Funktionen das Protein in der Zelle oder im Organismus übernimmt, wohingegen die biochemische Funktion die katalysierte Reaktion (z.B. Endopeptidase, Ligase) charakterisiert. Aus der Struktur eines Proteins lässt sich maximal die biochemische Funktion ableiten, d.h. sie wird im starkem Maße von der dreidimensionalen Gestalt des Proteins beeinflußt. Durch das Faltungsmuster werden katalytisch wichtige Reste in eine definierte räumliche Anordnung gebracht. So kann eine bestimmte chemische Reaktion katalysiert werden. Aus diesem Grund können neben Sequenzvergleichen auch Faltungsmusteranalysen Rückschlüsse über die Funktion eines Proteins liefern [Martin et al., 1998; Thornton et al., 1999; Nagano et al., 2002] (siehe auch Kapitel 2.2). Bestimmend für die Funktion eines Proteins sind aber vor allem die Anordnung und die physikochemischen Eigenschaften der Aminosäuren im aktiven Zentrum. Wie in Kapitel 5.1 beschrieben wurde, stellt diese Annahme die zentrale Hypothese des Cavbase Konzepts dar. Proteine mit ähnlicher Architektur im aktiven Zentrum sollten ähnliche Reaktionen katalysieren. In diesem Kapitel wird anhand einiger exemplarischer Beispiele gezeigt, wie Cavbase für die Annotierung und Funktionsanalyse von Proteinen genutzt werden kann. Ein Schwerpunkt liegt vor allem auf der Entdeckung von Ähnlichkeiten zwischen Proteinen, die mit Vergleichsmethoden, die auf Sequenz- oder Faltungsmethoden zurückgreifen, nicht festgestellt werden können. Im Abschnitt 4.1.1 werden als Beispiel für eine Ähnlichkeitsanalyse die Ergebnisse der Suche mit einer viralen Cysteinprotease vorgestellt. Des Weiteren werden in Abschnitt 4.1.2 zu erwartende und überraschende

Ähnlichkeiten von NAD(P)-bindenden Protein untersucht. Abschnitt 4.1.3 analysiert Verwandtschaften von Zink-bindenden Enzymen und in Abschnitt 4.2 wird die funktionelle Annotation von Proteinen mit bislang unbekannter Funktion mit Hilfe von Cavbase vorgestellt.

4.1.1 SARS-Coronavirus M^{pro}

Das SARS-Coronavirus (SARS-CoV) ist der Erreger des schweren akuten respiratorischen Syndroms (*Severe Acute Respiratory Syndrome, SARS*). Es besitzt eine Cysteinprotease (Main Proteinase, im folgenden M^{pro}), die wichtige Aufgaben bei der Virusreplikation übernimmt. In diesem Abschnitt wird die Ähnlichkeitssuche mit der SARS-CoV M^{pro} [Yang et al., 2003] und des humamen Coronavirus (HCoV) vorgestellt. Die HCoV M^{pro} ist der SARS-CoV-M^{pro} sehr ähnlich (Sequenzidentität 40%) [Anand et al., 2003] und die HCoV Kristallstruktur wurde ein halbes Jahr vor der SARS-CoV-M^{pro} Struktur gelöst. In Kapitel 6.1 wird das *de Novo-Design* von Inhibitoren, die für die SARS-CoV M^{pro} Subtaschen optimiert wurden, vorgestellt und die Krankheit SARS genauer beschrieben. Der Schwerpunkt in diesem Kapitel liegt auf der Analyse ähnlicher Bindetaschen und dem Auffinden von funktional verwandten Proteinen.

Cysteinproteasen besitzen in aller Regel eine Triade, allerdings sind auch Fälle bekannt, in denen nur eine Diade aus einem Cystein und einem Histidin vorliegt. Der Aufbau ihrer Bindetasche ist dem der Serinproteasen [Wallace et al., 1996; Branden and Tooze, 1999], die eine katalytische Triade bestehend aus einem Serin, Histidin und Aspartat besitzen, ähnlich. In Abbildung 4.1 ist das Faltungsmuster der HCoV M^{pro} dargestellt. Die HCoV M^{pro} und die SARS-CoV M^{pro} besitzen eine Serinprotease-ähnliche Faltung [Anand et al., 2002] bestehend aus zwei β-Faß-Domänen mit einer zusätzlichen α-helikalen Domäne. Die Bindetasche ist zwischen den beiden β-Faß-Domänen lokalisiert. Sie wird von Cavbase zuverlässig detektiert und umfasst alle katalytisch wichtigen Reste. Die Tasche wurde gegen einen Datensatz von 17345 Bindetaschen verglichen. In Abbildung 4.2 sind die 400 ersten Ränge der Ähnlichkeitssuche aufgezeigt. Cavbase findet auf den ersten 13 Rängen Bindetaschen von allen viralen Proteinen im Datensatz, wie Proteasen des HCoV, des Tabakätzvirus oder des Virus der Transmissiblen Gastroenteritis (TGEV). Auf die Gemeinsamkeiten und Unterschiede in der Bindetasche zweier M^{pro} Strukturen wird in Kapitel 6.2 näher eingegangen. Auf Rang 13 wird die 3C Proteasenstruktur des Rhinovirus mit gebundenen Inhibitor AG7088 (**1**, siehe

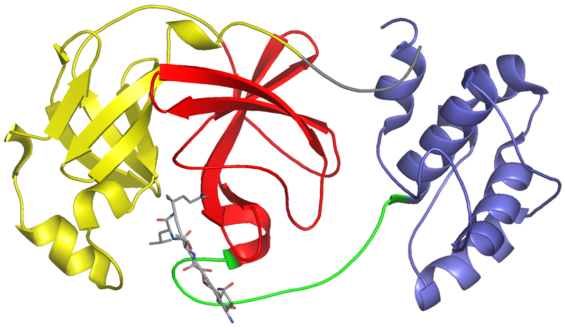


Abb. 4.1 Faltungsmuster der HCoV-SARS M^{pro} mit gebundenem peptidischen Inhibitor. Die HCoV-SARS M^{pro} Faltung besteht aus drei Subdomänen, die β-Faß-Domänen sind in gelb und rot, die α-helikale Domäne ist in blau dargestellt. Die Faltung dieser beiden Domänen ähnelt der Serinproteasefaltung (Chymotrypsin-ähnlich). Die Bindetasche ist zwischen den beiden β-Faß-Domänen lokalisiert.

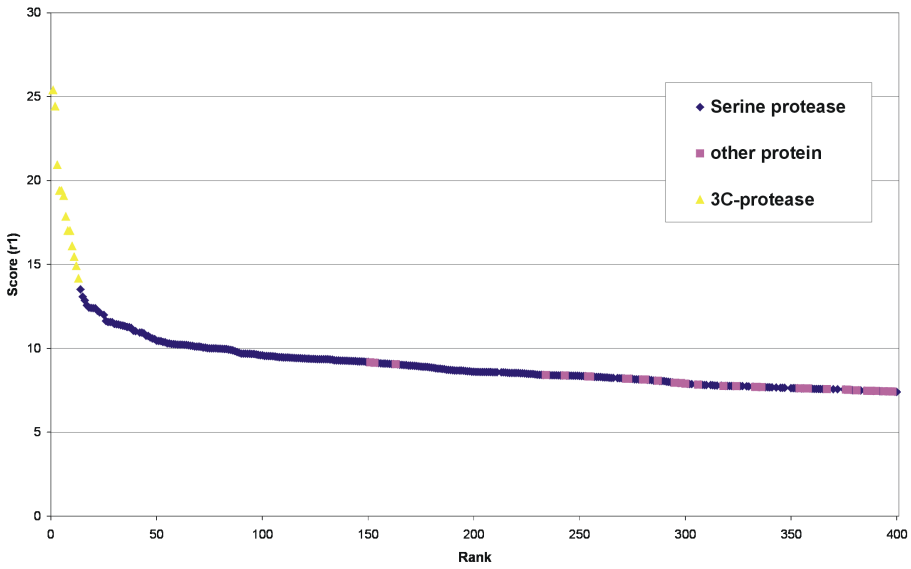
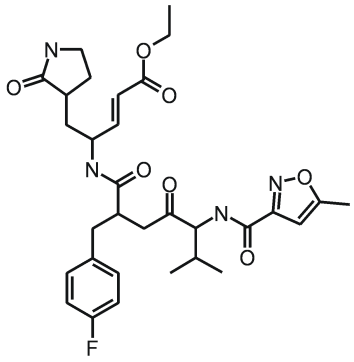


Abb. 4.2 Die 400 besten Ergebnisse der Vergleichsanalyse mit der SARS-CoV M^{pro} 1uk4 sortiert nach Bewertungsfunktion r1. In gelb sind zu 1uk4 verwandte Bindetaschen eingefärbt, Bindetaschen von Serinproteasen sind blau und Bindetaschen anderer Proteine in violett. Cavbase kann sehr gut Bindetaschen von stark verwandten Proteinfamilien identifizieren und vom Rest des Datensatzes separieren.

Abbildung 4.3) gefunden [Matthews et al., 1999; Dragovich et al., 1999]. Abbildung 4.4 zeigt die Überlagerung der ähnlichen Bereiche beider Enzyme. Der Inhibitor AG7088 befindet sich zur Zeit in klinischen Studien zur Therapie der Erkältung und wurde im Rahmen der SARS-Epidemie auf Aktivität an der SARS CoV M^{pro} getestet.



1

Abb. 4.3 Chemische Formel von AG7088. Eine Michael Akzeptorfunktion macht diese Verbindung zu einem irreversiblen Inhibitor der Rhinovirus 3C Protease. Er bildet eine kovalente Verknüpfung zum Schwefel des katalytischen Cysteins aus.

Mit Cavbase gelingt eine plausible Überlagerung beider Bindetaschen. Die katalytisch wichtigen Bereiche in der Bindetasche werden als zueinander ähnlich angesehen, wie zum Beispiel die katalytische Diade und auch der Bereich, in dem das Substrat über Hauptkettenatome gebunden wird. Das SARS Inhibitorpeptid und AG7088 überlagern sehr gut (RMSD Wert der über 17 als ähnlich angesehenen Pseudozentren ist 1.046Å), da beide ähnliche Interaktionen mit dem Protein ausbilden [Anand et al., 2003]. Die jeweiligen Seitenketten beider Inhibitoren adressieren die entsprechenden Subtaschen (für eine weitere Analyse siehe auch Kapitel 6).

Neben den zu erwartenden Ähnlichkeiten zu verwandten Kristallstrukturen findet man auf den nächsten Rängen sehr viele Bindetaschen von Serinproteasen. Cavbase zeigt Ähnlichkeiten in katalytisch wichtigen Bereichen beider Enzyme, wie in der katalytischen Diade/Triade, sowie in Bereichen der Substraterkennung durch die Atome der Peptidbindung der Hauptkette (*main chain substrate binding*) und im Bereich des sogenannten O^--Loch (*oxyaninon hole*). Hier entdeckt Cavbase den ähnlichen Aufbau funktionaler Bereiche, der diesen Proteasen gemeinsam ist. Beide Proteasen katalysieren dieselbe Reaktion: die Spaltung einer Peptidbindung. Als Beispiel für eine Überlagerung der SARS-CoV M^{pro} mit einer Serinprotease wurde die Bindetasche einer α-lytischen Protease ausgewählt, die als erste nicht verwandte Struktur auf Rang 14 gefunden wurde (siehe Abbildung 4.5). Vergleicht man die Ergebnisse der Bindetaschenähnlichkeiten mit anderen Vergleichsverfahren, wie Sequenzähnlichkeit und Faltungsmusterhomologie, erkennt man, daß in diesem Fall alle Ansätze ähnliche Ergebnisse liefern. Auch mit Hilfe von Sequenzvergleichen findet man auf den ersten Rängen Sequenzen verwandter, viraler Proteasestrukturen (Sequenzähnlichkeit zwischen HCoV und SARS CoV 40

%). Interessant ist aber die hohe Ähnlichkeit der SARS M^{pro} zu den Serinproteasen, die durch eine hohe Ähnlichkeit im Bereich der katalytischen Diade/Triade und der Hauptketten-Substratbindetstelle vermittelt wird. Das Wissen über ähnliche Bereiche in der Bindetasche lässt sich in diesem Fall am Besten zum Design von Subtaschen-optimierten Inhibitoren nutzen. Darauf wird in Kapitel 6.2 eingegangen.

4.1.2 NAD(P)-bindende Enzyme

In einer weiteren Untersuchung wurde die Bindetasche von NADH-bindenden Enzymen analysiert. Die Bindetasche der UDP-Galactose-4-Epimerase (PDB Code 1xel) wurde gegen einen Datensatz von 9145 Bindetaschen verglichen. Die Bindetasche der UDP-Galactose-4-Epimerase ist relativ groß und umfasst neben der NADH-Bindestelle ebenfalls noch die Bindestelle für UDP-Galactose/UDP-Glucose. Die UDP-Galactose-4-Epimerase katalysiert die Umwandlung von UDP-Galactose zu UDP-Glucose unter Reduktion von NAD^+, in dieser Struktur ist das Produkt UDP-Glucose gebunden.

Auf den ersten Rängen wurden Bindetaschen von anderen Enzymen mit gebundenem NAD(P)H gefunden. Die Kofaktoren überlagern sehr gut, die an der Bindung beteiligten Aminosäuren sind sich sehr ähnlich (siehe Abbildung 4.6). Auf den weiteren Plätzen folgen Bindetaschen, die zum NADH verwandten Kofaktoren (z.B. SAM und FAD) gebunden haben. Interessanterweise wird auf Rang 149 die Bindetasche einer Glucoseoxidase (1gal) entdeckt, die zu 1xel keine Sequenz- und Faltungsähnlichkeiten zeigt (siehe Tabelle 4.1). Die UDP-Galctose-4-Epimerase besitzt eine Rossmann-Faltung, die Glucoseoxidase weist eine FAD/NAD(P)-Bindedomäne auf. Die Aminosäuren, die an der Bindung des Kofaktors beteiligt sind, sind komplett verschieden. Dennoch exponieren sie ähnliche physikochemische Eigenschaften in die Tasche und werden daher von Cavbase als ähnlich erkannt (siehe Abbildung 4.8).

4.1.3 Verwandtschaftsbeziehungen zwischen NEP und Thermolysin

Die Neutrale Endopeptidase (neutral endopeptidase, NEP) und Thermolysin gehren beide zur großen Familie der Metalloproteasen, die ein Zink-Ion im aktiven Zentrum besitzen. Beide Enzyme zeigen untereinander nur 16% Sequenzidentität. NEP besteht

Tab. 4.1 Als äquivalent erkannte Pseudozentren und Aminosäuren in den Bindetaschen der UDP-Galactose-4-Epimerase und der Glucose Oxidase.

UDP-Galactose-4-Epimerase (1xel.1)			**Glucose Oxidase** (1gal.1)		
Typ des Pseudo-zentrums	äquivalente Aminosäure[a]		Typ des Pseudo-zentrums	äquivalente Aminosäure[a]	
Donor	I 12	p	Donor	L 29	p
Pi	L 30	p	Pi	I 49	p
Akzeptor	L 30	p	Akzeptor	I 49	p
Akzeptor	D 31	s	Akzeptor	E 50	s
Akzeptor	D 31	s	Akzeptor	E 50	s
Donor	N 32	p	Donor	S 51	p
Donor	N 32	s	Donor-Akzeptor	S 51	s
Donor	N 35	p	Donor-Akzeptor	H 78	s
Donor-Akzeptor	S 36	s	Donor	G 99	p
Akzeptor	G 57	p	Akzeptor	Q 248	p
Donor	I 59	p	Donor	V 250	p
Akzeptor	G 82	p	Akzeptor	A 289	p
Donor	K 84	p	Donor	A 292	p

[a] Ein-Buchstabencode der Aminosäure, Aminosäurenummer und Ursprung des Pseudozentrums: aus der Seitenkette (s) oder aus der Peptidbindung (p).

aus drei Domänen, mit einer Thermolysin-ähnlichen katalytischen Domäne. Die Bindetasche, die das aktive Zentrum von NEP (PDB Code 1dmt [Oefner et al., 2000]) umfasst, wurde gegen 9145 Bindetaschen verglichen. Auf den ersten Rängen werden viele Strukturen verschiedener Metalloproteasen gefunden. Eine Thermolysinstruktur wird auf Rang eins angetroffen, die katalytischen Reste und gebundenen Inhibitoren überlagern sehr gut. Auf den nächsten Rängen folgen weitere Beispiele der Metalloproteasenfamilie mit Zink im aktiven Zentrum (Collagenasen, Stromelysin, etc.). Ab Rang 200 werden dann auch Proteine gefunden, die andere Metallionen im aktiven Zentrum gebunden haben.

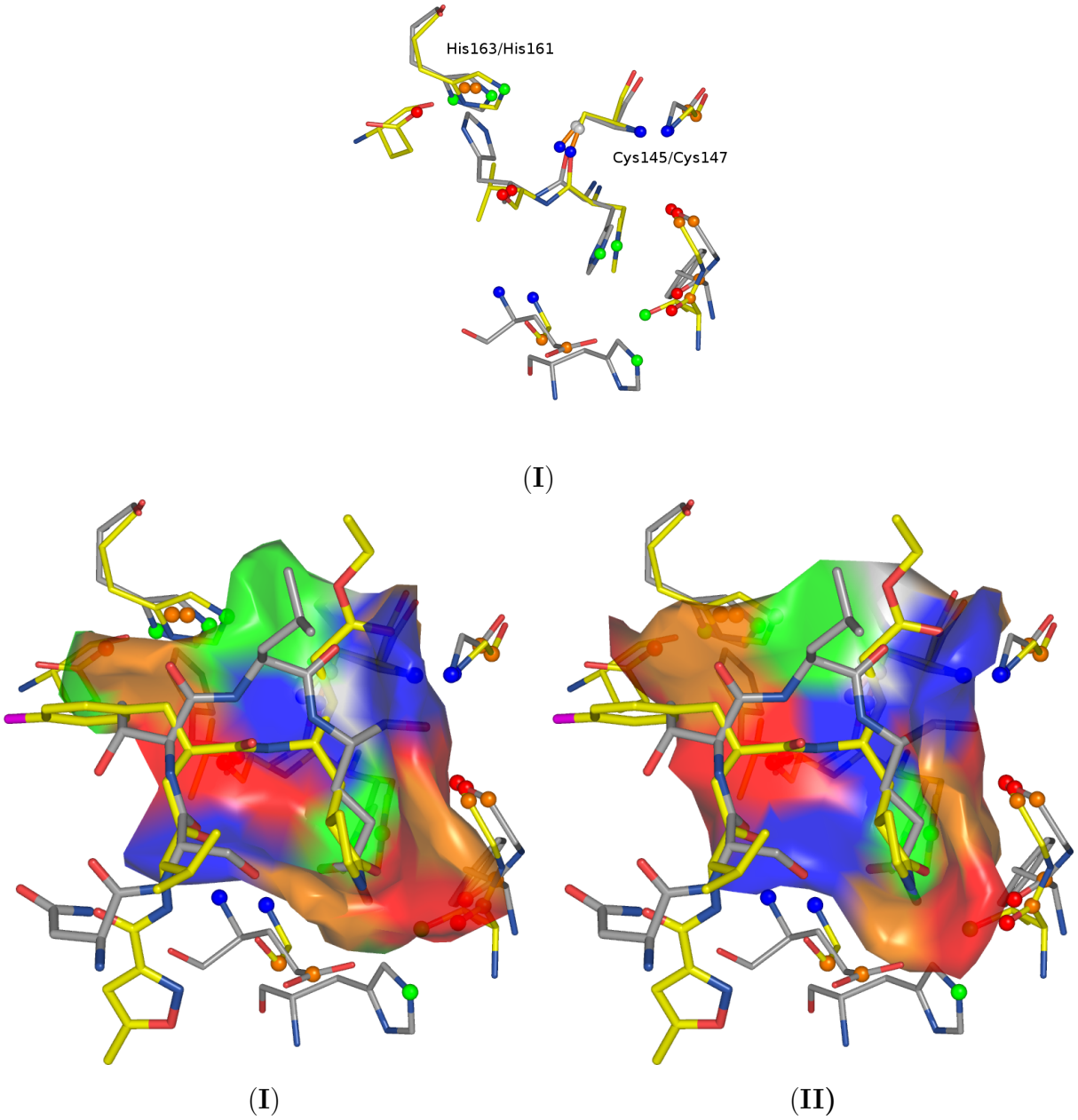


Abb. 4.4 Überlagerung der als ähnlich erkannten Bereiche der Bindetaschen der SARS M^{pro} (PDB Code 1uk4) mit der Rhinovirus 3C Protease (PDB Code 1cqq). In (I) ist die Überlagerung der vom Vergleichsalgorithmus als ähnlich erkannten Aminosäuren und Liganden gezeigt, SARS M^{pro} (Kohlenstoffe in grau), Rhinovirus 3C Protease (Kohlenstoffe in gelb). Man erkennt, daß die katalytischen Aminosäuren gut überlagern, auch werden die Inhibitoren gut zur Übereinstimmung gebracht (II und III). Die jeweiligen Subtaschen werden von den entsprechenden Resten adressiert. Dargestellt sind die vom Vergleichsalgorithmus als ähnlich erkannten Aminosäuren, Liganden und Oberflächenbereiche (in (II) für die SARS M^{pro} Protease, in (II) für die Rhinovirus 3C Protease); Farbcodierung wie in (I).

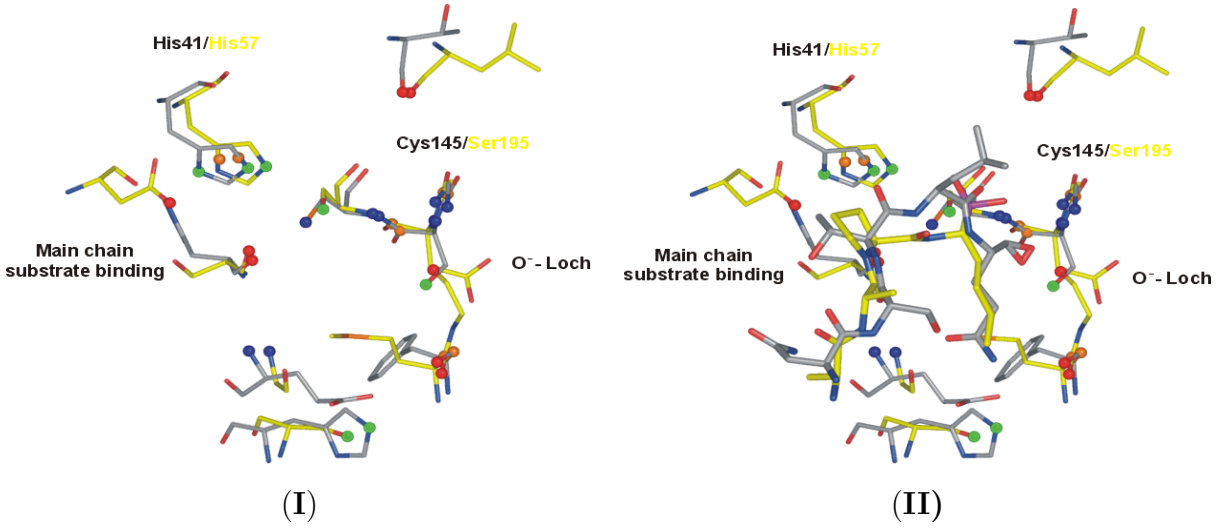


Abb. 4.5 Überlagerung der als ähnlich erkannten Bereiche der Bindetaschen der SARS M^{pro} (PDB Code 1uk4) und der α-lytischen Protease (PDB Code 6lpr). In (I) ist die Überlagerung der vom Vergleichsalgorithmus als ähnlich erkannten Aminosäuren und Pseudozentren gezeigt. Funktionell wichtige Bereiche, wie die katalytische Diade, die *main chain substrate binding*) und das sogenannte O^--Loch (*oxyaninon hole*) werden als ähnlich erkannt. Die SARS-CoV M^{pro} ist mit Kohlenstoffen in grau, die α-lytische Protease mit Kohlenstoffen in gelb dargestellt. In (II) sind zusätzlich die gebundenen Liganden gezeigt. Farbcodierung wie in (I).

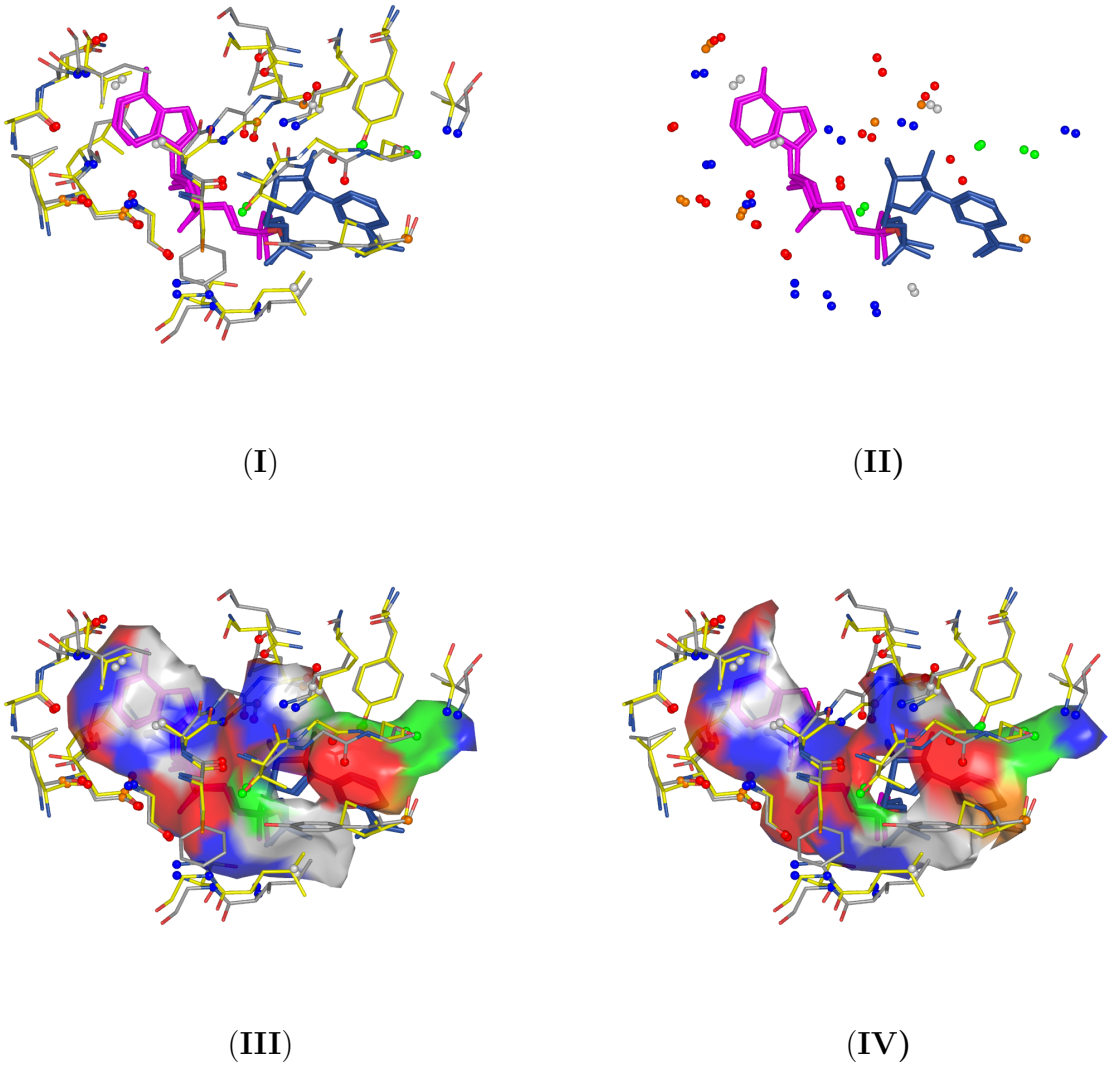


Abb. 4.6 Vom Vergleichsalgorithmus als ähnlich erkannte Bereiche in den Bindetaschen der UDP-Galactose-4-Epimerase (PDB Code 1xel) und einer Acyl-CoA-Dehydrogenase (PDB Code 1e6w). In (I) sind als ähnlich erkannten Aminosäuren und Liganden gezeigt, die UDP-Galactose-4-Epimerase mit Kohlenstoffen in grau, die Acetyl-CoA-Dehydrogenase mit Kohlenstoff-Atomen in gelb. Man erkennt, daß die Aminosäuren, die sich an der Kofaktor-Bindung beteiligen, identisch sind. Die Kofaktorbindestelle ist strukturell weitgehend erhalten und zeigt keine große strukturelle Abweichung. (II) zeigt die beiden Kofaktoren und die Pseudozentren, die in beiden Taschen als ähnlich gefunden wurden. In (III) und (IV) sind zusätzlich die vom Vergleichsalgorithmus als ähnlich erkannten Aminosäuren und jeweils die Oberflächenbereiche von der UDP-Galactose-4-Epimerase (III) und der Acyl-CoA-Dehydrogenase (IV) dargestellt.

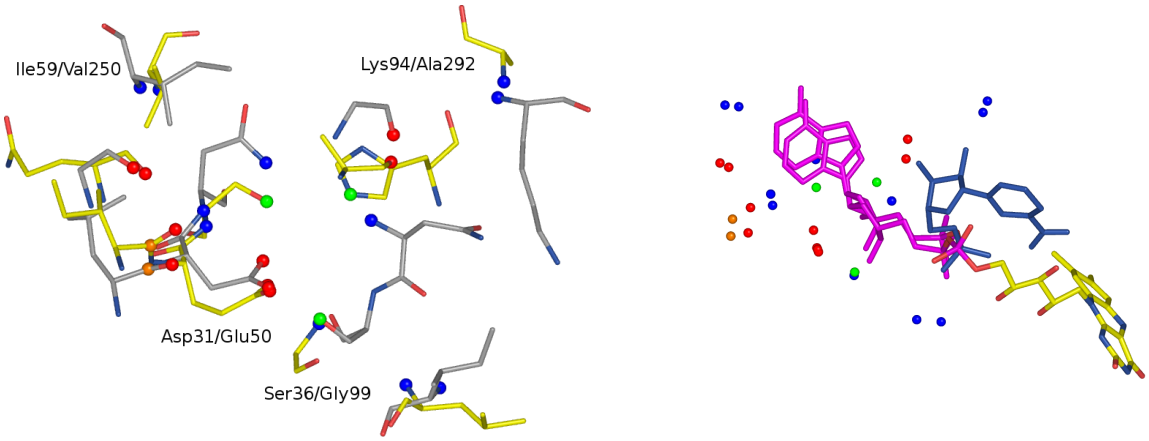


Abb. 4.7 Vom Vergleichsalgorithmus als ähnlich erkannte Bereiche in den Bindetaschen der UDP-Galactose-4-Epimerase (PDB Code 1xel) und der Glucose Oxidase (PDB Code 1gal). In (I) sind als ähnlich erkannte Aminosäuren und Pseudozentren gezeigt, die UDP-Galactose-4-Epimerase mit Kohlenstoffen in grau, die Glucose Oxidase mit Kohlenstoff-Atomen in gelb. Die Aminosäuren, die sich an der Bindung des Kofaktors beteiligen, sind komplett unterschiedlich. Sie bilden aber ein ähnliches Interaktionsmuster aus, was von Cavbase erkannt wird. In (II) sind zusätzlich die gebundenen Liganden dargestellt.

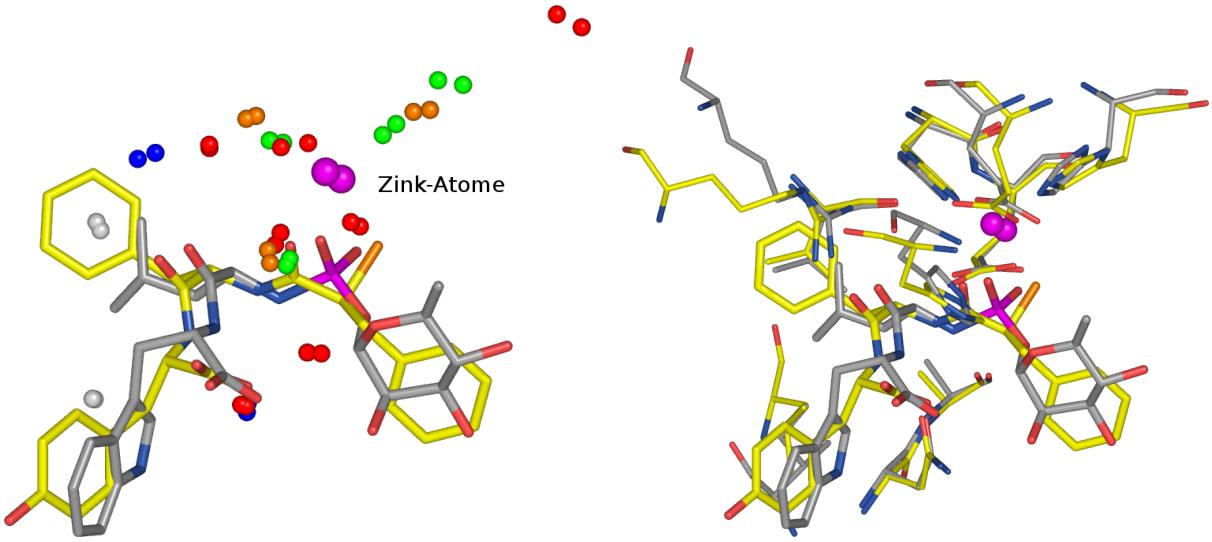


Abb. 4.8 Vom Vergleichsalgorithmus als ähnlich erkannte Bereiche in den Bindetaschen der NEP (PDB Code 1dmt) und Thermolysin (PDB Code 1qf0). In (I) sind als ähnlich erkannte Pseudozentren und Liganden gezeigt, wobei die Kohlenstoffe in NEP grau und in Thermolysin in gelb wiedergegeben sind. Man erkennt, daß die katalytisch wichtigen Aminosäuren, die an der Inhibitor-Bindung beteiligt sind sehr gut überlagern, auch die Reste, die die verschiedenen Subtaschen adressieren überlagern gut. In (II) sind nur die entsprechenden Aminosäuren und Liganden dargestellt, Farbcodierung wie in (I).

4.2 Proteinstrukturen mit unbekannter Funktion als Testfall für die Annotation von Proteinen

In breit angelegten *structural genomics* Initiativen (MCSG [MCSG], SPINE [SPI], SGC [SGC], PSI [PSI]) wird seit Ende der neunziger Jahre versucht, eine große Anzahl neuer Proteinstrukturen zu lösen [Brenner and Levitt, 2000; Burley, 2000; Thornton, 2001; Mittl and Grutter, 2001; Burley and Bonanno, 2003; Yee et al., 2003]. Ein Schwerpunkt dieser Initiativen bildet die Aufklärung von Proteinstrukturen, für die noch keine Faltungs-homologen Proteine bekannt sind bzw. von Proteinen und Proteinkomplexen mit unmittelbarer medizinischer Relevanz. Diese Arbeiten sollen sich außerdem durch einen hohen Grad an Automatisierung und einen hohen Durchsatz auszeichnen [Watson et al., 2003]. Aus diesem Grund wurden in den letzten Jahren viele Techniken zur Hochdurchsatz-Aufklärung von Proteinstrukturen entwickelt. In diesen Punkten unterscheiden sich die *structural genomics* Initiativen von klassischer Strukturaufklärung, bei der zumeist einzelne Proteine sequentiell gelöst werden und in aller Regel eine Vorstellung über die Funktion des Proteins zuvor vorhanden war. Proteine, deren Funktion zum Zeitpunkt der Strukturaufklärung noch nicht bekannt sind, werden als *'hypothetical proteins'* bezeichnet [Eisenstein et al., 2000]. Ende 2003 waren in der PDB über 300 *'hypothetical proteins'* Strukturen abgelegt, wobei man für knapp 50% Prozent dieser Proteine im Nachhinein die Funktion zuweisen konnte [Yee et al., 2003]. Gerade diese Strukturen stellen interessante Testfälle für die Methoden dar, die die Funktion eines Proteins aus dessen Struktur ableiten. Tabelle 4.2 zeigt eine Reihe dieser *'hypothetical proteins'*, die zur Ähnlichkeitssuche mit Cavbase benutzt wurden. Alle diese Strukturen wurden gegen einen Datensatz von 17337 Bindetaschen verglichen und auf Ähnlichkeiten zu bekannten Strukturen untersucht. Bei einem Bindetaschenvergleich von *'hypothetical proteins'* müssen einige wichtige Punkte beachtet werden. Da die Funktion dieser Proteine unbekannt ist, ist keine Informationen über funktionelle Aminosäuren vorhanden, vor allem ist die genaue Lage der Bindetasche nicht gegeben. Insbesondere haben aber Lage, Größe und Eigenschaften der automatisch detektierten Bindetasche starken Einfluß auf die Ergebnisse der Ähnlichkeitssuche. Wenn die katalytische Bindetasche in einem flachen Bereich auf der Proteinoberfläche liegt, wird sie mit dem Bindetaschendetektionsalgorithmus nicht gefunden. Problematisch sind auch solche Fälle, in denen Cavbase nur einen Teil der katalytisch wichtigen Aminosäuren findet.

Aufschluß über die Funktion eines Proteins kann man zum einen durch Sequenz- und/oder Faltungsvergleiche, aber auch durch biochemische Experimente erhalten, wie an einigen Beispielen von Proteinen mit unbekannter Struktur in Kapitel 4.2.1 gezeigt wird. Dabei stellt sich die Frage, ob man diese Annotationen auch strukturell erklären und mit Cavbase nachvollziehen kann. In Abschnitt 4.2.2 wird eine ausführliche Ähnlichkeitsanalyse für ein ATP-bindendes Protein vorgestellt und die Ergebnisse der Ähnlichkeitssuche diskutiert.

4.2.1 Ähnlichkeitsanalysen mit *'hypothetical proteins'*

Ein Pilotprojekt zur Aufklärung von mehreren hundert Proteinstrukturen aus einem thermophilen Organismus hat bis 2003 zur Veröffentlichung von 36 Strukturen geführt [Yee et al., 2003], ein Großteil der Strukturen ist in Tabelle 4.2 aufgelistet. Anhand dieser Proteinstrukturen wird die Ähnlichkeitssuche mit Proteinstrukturen unbekannter Funktion vorgestellt und am Beispiel einiger Vergleiche werden die Möglichkeiten und Grenzen der Funktionsannotation mit Cavbase diskutiert.

Man kann die untersuchten Proteine mit unbekannter Funktion nach der erzielten Annotation in drei Gruppen einteilen. Die erste Gruppe bilden Proteine mit unbekannter Funktion, für die Bindetaschen gefunden werden, die eine große Ähnlichkeit zur Suchtasche aufweisen und somit Rückschlüsse auf die Funktion des Proteins erlauben. Zusätzlich können gebundene Liganden bei der Funktionszuweisung helfen. Bei einigen der im Nachfolgendem vorgestellten Strukturen wurde die Funktion des Proteins neben der Sequenzähnlichkeit zusätzlich noch durch strukturelle Vergleiche oder biochemische Experimente weiter annotiert. Immer dann, wenn eine ausreichende Menge ähnlicher Proteinbindetaschen vorliegt, ist Cavbase in der Lage, verwandte Proteine zu identifizieren. In die zweite Gruppe fallen Proteine, für die Bindetaschen mit ähnlichen lokalen Bereichen gefunden werden, bei denen aber die gesamte Tasche keine Ähnlichkeit mit der Suchtasche aufweist. Wenn in den als ähnlich erkannten Bereichen zusätzlich Liganden gebunden sind, erlauben diese Rückschluß auf die Funktion des Proteins. Die dritte Gruppe bilden Proteine, für die keine signifikanten Ähnlichkeiten gefunden werden. Diese Ähnlichkeiten, die zwischen den Bindetaschen entdeckt werden, sind hier auf kleine, verstreute Bereiche in der Tasche beschränkt, zusätzlich kann über die gebundenen Liganden keine weitere Information über mögliche Ähnlichkeiten gewonnen werden. Zwei beliebige Bindetaschen besitzen in fast jedem Fall einen gewissen Grad

an Ähnlichkeit (Muster bestehend aus vier Pseudozentren werden fast immer gefunden). In diesen Fällen ist eine Funktionsannotierung mit Cavbase nicht möglich. Mit zunehmender Zahl an gelösten Proteinstrukturen wird man in Zukunft immer häufiger in der Lage sein, eine strukturell ähnliche Bindetasche zu detektieren. Funktionszuweisungen basierend auf Ähnlichkeiten in der Bindetasche werden dann mit größerer Wahrscheinlichkeit Erfolg haben.

Als Beispiel für ein Protein, für das keine verwandte Bindetasche detektiert werden kann, dient die RNA Polymerase Untereinheit 10 (RPB10, PDB Code 1ef4 [Mackereth et al., 2000]). Hierbei handelt es sich um ein relativ kleines Zink-bindendes Protein bestehend aus 55-80 Aminosäuren (je nach Ursprungsorganismus), welches Bestandteil von RNA Polymerasen (*multisubunit RNA polymerases (RNAPs)* ist. Es hat - gerade wegen seiner Größe - keine direkte katalytische Funktion, sondern vermittelt seine funktionelle Bedeutung über eine strukturelle Interaktion mit anderen Untereinheiten der RNAPs. Auf der Oberfläche befinden sich deshalb auch keine (funktionellen) Vertiefungen (siehe Abbildung 4.9-I). Deshalb ist dieses Protein für eine Ähnlichkeitsanalyse mit Cavbase nicht geeignet.

Im Folgenden werden Beispiele für eine Ähnlichkeitssuche mit Proteinen vorgestellt, für die Cavbase keine signifikanten Ähnlichkeiten zu bekannten Proteinbindetaschen findet. Eine Ähnlichkeitsanalyse kann hier aber Ideen für Liganden oder Ligandenfragmente liefern, die eine Affinität zu diesem Protein aufweisen könnten. Dieses Wissen könnte zur besseren Planung von biochemischen Experimenten zur Funktionszuweisung verwendet werden.

Die Phosphoribosylformylglycinamidin (FGAM) Synthetase ist an der *de novo* Biosynthese von Purinbausteinen beteiligt [Batra et al., 2002]. In Eukaryonten und einigen Bakterien besteht die FGAM Synthase aus einem großen Protein mit zwei Domänen: einer ATPase-Domäne und einer Glutamin-Bindedomäne. In Archebakterien und anderen Bakterien wird die FGAM Synthase durch zwei verschiedene Gene kodiert und agiert als Protein, welches aus mehreren Untereinheiten besteht. In *Bacillus subtilis* ist das *purL* Gen homolog zu der ATPase Domäne und das *purQ* Gen zur Glutamin-Bindedomäne. Auf dem gleichen Operon liegt ein weiteres Gen, das für das *purS* Protein kodiert. Ohne das Vorhandensein des *purS* Proteins kann das Bakterium aufgrund fehlerhafter FGAM-Synthase Aktivität kein Purin synthetisieren. Mth0169 (PDB Code 1gtd) zeigt 29 % Sequenzidentität zu dem *purS* Protein und Batra et al. gehen davon aus, daß es sich bei Mth0169 um ein orthologes Protein zu *purS* handeln könn-

te. Die genaue biochemische Funktion ist aber noch ungeklärt. Für Mth0169 werden zwei Bindetaschen gefunden. Die beiden Ähnlichkeitsanalysen ergeben jedoch keine signifikanten Ähnlichkeiten zu schon bekannten Proteinen. Es werden aber auch Bindetaschen gefunden, die lokale Ähnlichkeiten zu Mth0169 aufweisen und die Liganden oder Ligandfragmente in mit der Mth0169 Bindetasche verwandten Bereichen gebunden haben.

Die Kristallstruktur von Mth0938 (PDB Code 1ihn) zeigt ein neues Faltungsmuster [Das et al., 2001] und hat nach der SCOP-Datenbank keine strukturellen Nachbarn. In diesem Fall detektiert Cavbase zwei Bindetaschen auf der Oberfläche des Proteins. Eine Bindetasche umfasst Aminosäuren in einem polaren Bereich des Proteins, die in einer Analyse der elektrostatischen Eigenschaften der Aminosäuren aufgefallen waren [Das et al., 2001]. Die Cavbase Ähnlichkeitssuche findet in diesem Bereich der Bindetasche Ähnlichkeiten zu anderen Bindetaschen, die Liganden mit Phosphatgruppen gebunden haben. Möglicherweise zeigen Liganden mit einer negativ geladenen Gruppe wie die Phosphatgruppe Affinität zu Mth0938.

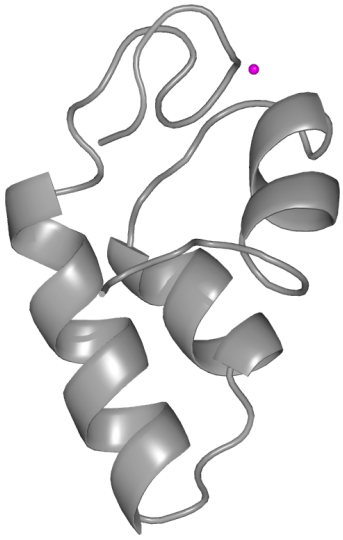


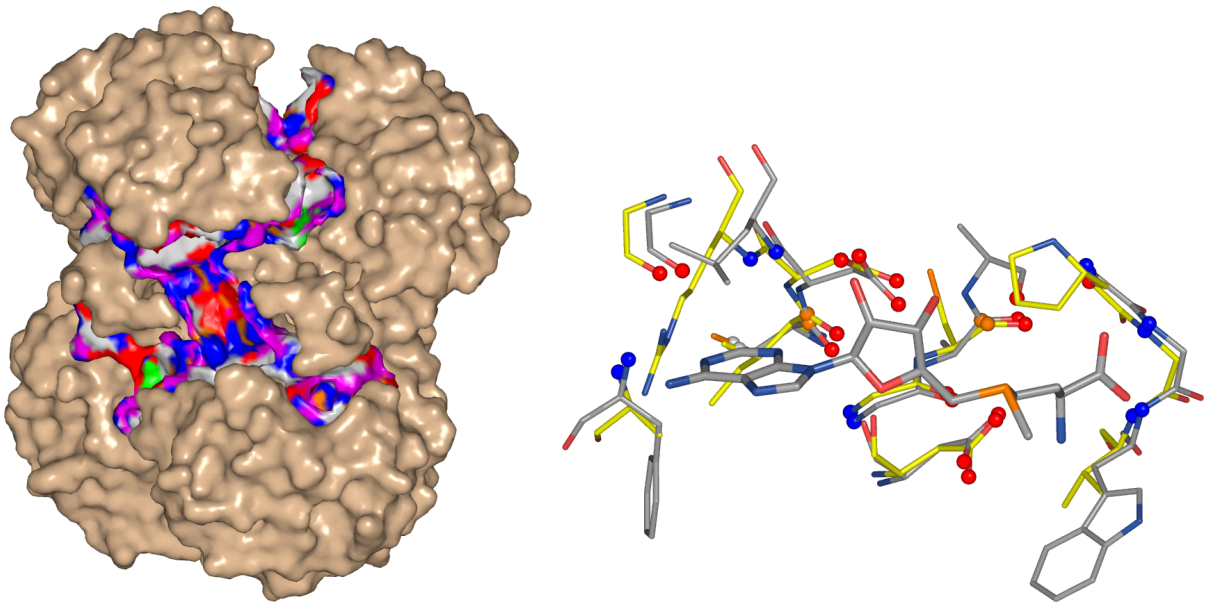
Abb. 4.9 Beispiel für eine Ähnlichkeitssuche mit Cavbase, in denen die Funktionsannotierung nicht erfolgreich war. Die Abbildung zeigt das Faltungsmuster der RNA Polymerase Untereinheit 10 (PDB Code 1ef4) mit einem gebundenen Zinkatom. Das Protein zeigt auf der Proteinoberfläche keine Vertiefungen, die als mögliche Bindestellen erkannt werden könnten.

Im Folgenden werden Beispiele für eine erfolgreiche Annotation von Proteinstrukturen mit Cavbase vorgestellt. Eine Analyse mit Proteinen unbekannter Funktion wird dann als erfolgreich angesehen, wenn Cavbase in der Lage ist, Bindetaschen zu finden, die

sehr große Gemeinsamkeiten zur Suchtasche aufweisen. Ein weiteres Kriterium für eine erfolgreiche Annotation stellt das sinnvolle Überlagern von Bindestellen für verwandte Liganden in zwei Proteinen dar.

Die CbiT und CbiE Enzyme sind in die Biosynthese von Vitamin B12 involviert und katalysieren zwei Methylierungen und eine Decarboxlierung an einer Precorrin Zwischenstufe (eine zyklische Tetrapyrrol-Verbindung) [Keller et al., 2002]. In einigen Organismen bilden beide Enzyme ein einziges Protein (CobL), das beide Reaktionen katalysiert. Man geht aufgrund der Sequenzhomologie von CbiE zu Precorrin Methyltransferasen davon aus, daß CbiT die Decarboxylierung katalysiert. MT0146 ist ein Sequenzhomologes zu CbiT, die Kristallstruktur zeigt aber erstaunlicherweise Faltungsähnlichkeiten zu S-Adenosyl-Methionin abhängigen Methyltransferasen. Dies deutet darauf hin, daß es sich bei Mt0146 um die vermutete Precorrin Methyltransferase handeln könnte. Mt0146 wäre das erste Enzyme mit Precorrin Methyltransferase Aktivität, das nicht die typische Precorrin Methyltransferase Faltung aufweist. Die Kristallstruktur von MT0146 Precorrin 8W Decarboxylase (PDB Code 1f38) zeigt, daß das Protein aus vier homologen Untereinheiten aufgebaut ist. Cavbase detektiert 10 Bindetaschen auf der Proteinoberfläche (siehe Abbildung 4.10-I). Einige der Bindetaschen werden an der Schnittstelle zwischen den einzelnen Proteindomänen entdeckt. Hat man über die ungefähre Lage der Bindestelle und katalytisch wichtige Reste keine Information, ist man auf den blinden Vergleich aller Bindetaschen angewiesen. Bei der Mehrzahl der in dieser Arbeit betrachteten Fälle ist die katalytische Tasche zugleich auch eine der größten Vertiefungen auf der Proteinoberfläche und Cavbase ist treffsicher in der Lage, diese zu detektieren. Im vorliegenden Fall zeigt der Vergleich der größten Bindetasche signifikant höhere Ähnlichkeiten zu anderen Bindetaschen aus dem Testdatensatz, die sich von den Ähnlichkeitswerten der anderen Taschen absetzen. So findet Cavbase die Bindetasche eines Hitzeschocksproteins (PDB Code 1eiz) FtsJ, das eine Methyltransferase Faltung aufweist und S-Adenosyl-Methionin als Kofaktor gebunden hat. Die Kofaktorbindestelle überlagert sehr gut mit der möglichen Bindestelle von S-Adenosyl-Methionin in Mth0146 (siehe Abbildung 4.10-II).

Die Kristallstruktur für Mth0152 (PDB Code 1eje) konnte mit einem gebundenen Flavin-Mononukleotid gelöst werden [Christendat et al., 2002]. Auf der Proteinoberfläche werden drei Bindetaschen gefunden, wovon eine das Flavin-Molekül gebunden hat. Für diese Bindetaschen werden sehr große Ähnlichkeiten mit Proteinen, die Flavin-artige Kofaktoren gebunden haben, gefunden (nicht dargestellt).



(**I**) (**II**)

Abb. 4.10 Ergebnisse der Ähnlichkeitsanalyse mit einer Precorrin 8W Decarboylase (PDB Code 1f38). In (I) sind vier der zehn Bindetaschen der Precorrin 8W Decarboxylase (PDB Code 1f38) gezeigt. Das Protein besteht aus vier Untereinheiten, an deren Schnittstelle ebenfalls Bindetaschen detektiert werden. Ein Problem bei Proteinen mit mehreren Bindetaschen ist die Auswahl der geeigneten Suchtasche, wenn keine Informationen über funktional wichtige Bereiche des Proteins vorliegen. Vergleicht man hier alle Taschen, setzt sich die katalytische Bindetasche von den anderen Taschen durch signifikant höhere Ähnlichkeitswerte zu anderen Funktions-zugewiesenen Taschen ab. (II) zeigt die Bindetasche der Precorrin 8W Decarboxylase (Kohlenstoffe in grau), die große Ähnlichkeiten mit einer RNA Methyltransferase FtsJ (PDB Code 1eiz) mit gebunden S-Adenosyl-Methionin (Kohlenstoffe in grau) hat.

Mth1790 ist eine Deoxythymidin Diphosphat (dTDP)-4-Keto-6-Deoxy-D-Hexulose-3, 5-Epimerase (RmlC), die in Bakterien als eines von vier Enzymen an der Biosynthese des Zellwandbestandteils dTDB-L-Rhamnose beteiligt ist (PDB Code 1epz) [Christendat et al., 2000]. dTDB ist außerdem ein Substrat für DNA-Polymerasen und in die Synthese von DNA involviert. Inhibition des Enzyms und die damit verbundene Störung des bakteriellen Zellwandaufbaus macht RmlC zu einem möglichen antibakteriellen Target. Cavbase detektiert für dieses Protein eine Bindetasche. Auf den ersten 11 Rängen werden Bindetaschen anderer Epimerasen gefunden (siehe Abbildung 4.11-I). Interessanterweise werden unter den ersten 50 Rängen 14 Trypsin-artige Serinproteasen gefunden, die alle im Bereich der *main chain substrate binding* und der S1-Tasche Ähnlichkeiten zu Mth1790 aufweisen (siehe Abbildung 4.11-II). Dieses Wissen über ähnliche

Oberflächenbereiche in beiden Bindetaschen könnte für das Design von antibiotischen Arzneistoffen nützlich sein.

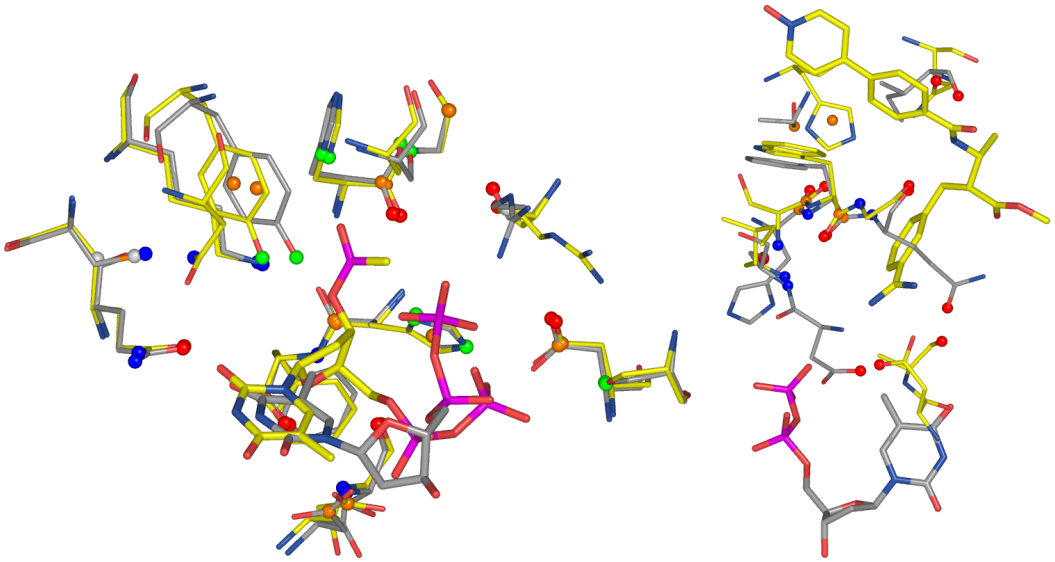


Abb. 4.11 Ergebnisse der Ähnlichkeitsanalyse eines Proteins, das an der Biosynthese bakterieller Zellwandbestandteile (PDB Code 1epz) beteiligt ist. In (I) sind die ähnlichen Bereiche in der Bindetasche zweier RmlC-Epimerasen gezeigt (PDB Code 1epz und 1dzt), große Ähnlichkeiten in der Bindetasche lassen auf dieselbe katalysierte Reaktion schließen. (II) zeigt die ähnlichen Bereiche der RmlC-Epimerase mit der Serinprotease Faktor Xa. Ähnlichkeiten werden hier im Bereich der S1-Tasche und der *main chain substrate binding* des Serinprotease gefunden.

Mth0129 kodiert für eine Nicotinamid Mononukleotid Adenylyltransferase (NMNATase, PDB Code 1ej2) [Saridakis et al., 2001]. Das Enzym ist an der Synthese von NAD beteiligt. Zur Zeit der Strukturaufklärung wurde die Sequenz als 'Protein unbekannter Funktion' annotiert, obwohl Verwandtschaften zu NMNATasen anderer Organismen bestanden. Vergleicht man die Bindetasche mit Cavbase, findet man auf den ersten Plätzen Bindetaschen von Proteinen mit verwandter Faltung, die ebenfalls NAD oder ähnliche Kofaktoren gebunden haben. Beispielsweise findet man auf Rang 7 die Bindetasche einer Cytidylyltransferase (PDB Code 1coz) mit gebundenem Cytidin-5-Triphosphat [Weber et al., 1999]. Das Cytidin überlagert sehr gut mit dem Adenosin des NADs und zeigt die ähnlichen Eigenschaften der Bindetasche im Bereich der Kofaktorbindestelle (nicht dargestellt).

Die Kristallstruktur von Mth1020 gehört zur SCOP-Faltungsklasse der N-terminalen Nucleophil (Ntn) Hydrolasen. Proteine dieser Faltungsklasse katalysieren die Hydro-

lyse von Amidbindungen und weisen denselben Reaktionsmechanismus auf [Saridakis et al., 2002]. Die Seitenkette von aminoterminalen Resten (Serin, Threonin, Cystein) dient als Nucleophil, und greift den Carbonylkohlenstoff an [Brannigan et al., 1995]. Das Nucleophil wird durch einen autokatalytischen Endoproteolyseschritt freigesetzt. Das Faltungsmuster und die Lokalisation der Bindetasche von Mth1020 und Ntn-Hydrolasen sind sehr ähnlich, es fehlt aber die N-terminale nucleophile Aminosäure in der Bindetasche - an dieser Stellte besitzt Mth1020 ein Arginin. Die Bindetasche, wie sie von Cavbase detektiert wird, umfasst die potentiell katalytischen Reste (siehe Abbildung 4.12-I). Eine Cavbase Ähnlichkeitsanalyse findet auf den ersten Rängen Bindetaschen von Proteinen, die eine Thymidylat Synthase/dCMP Hydroxymethylase oder eine Flavodoxin-ähnliche Faltung aufweisen. Auf dem ersten Platz wird die Bindetasche einer Deoxycytidylate Hydroxymethylase (PDB Code 1be5) gefunden (siehe Abbildung 4.12-II). Auffällig ist, daß unter den ersten Rängen viele Bindetaschen sind, die Phosphatgruppen gebunden haben, was auf eine mögliche Bindungsaffinität von Liganden mit Phosphatgruppen hindeuten könnte (siehe Abbildung 4.12-III und IV).

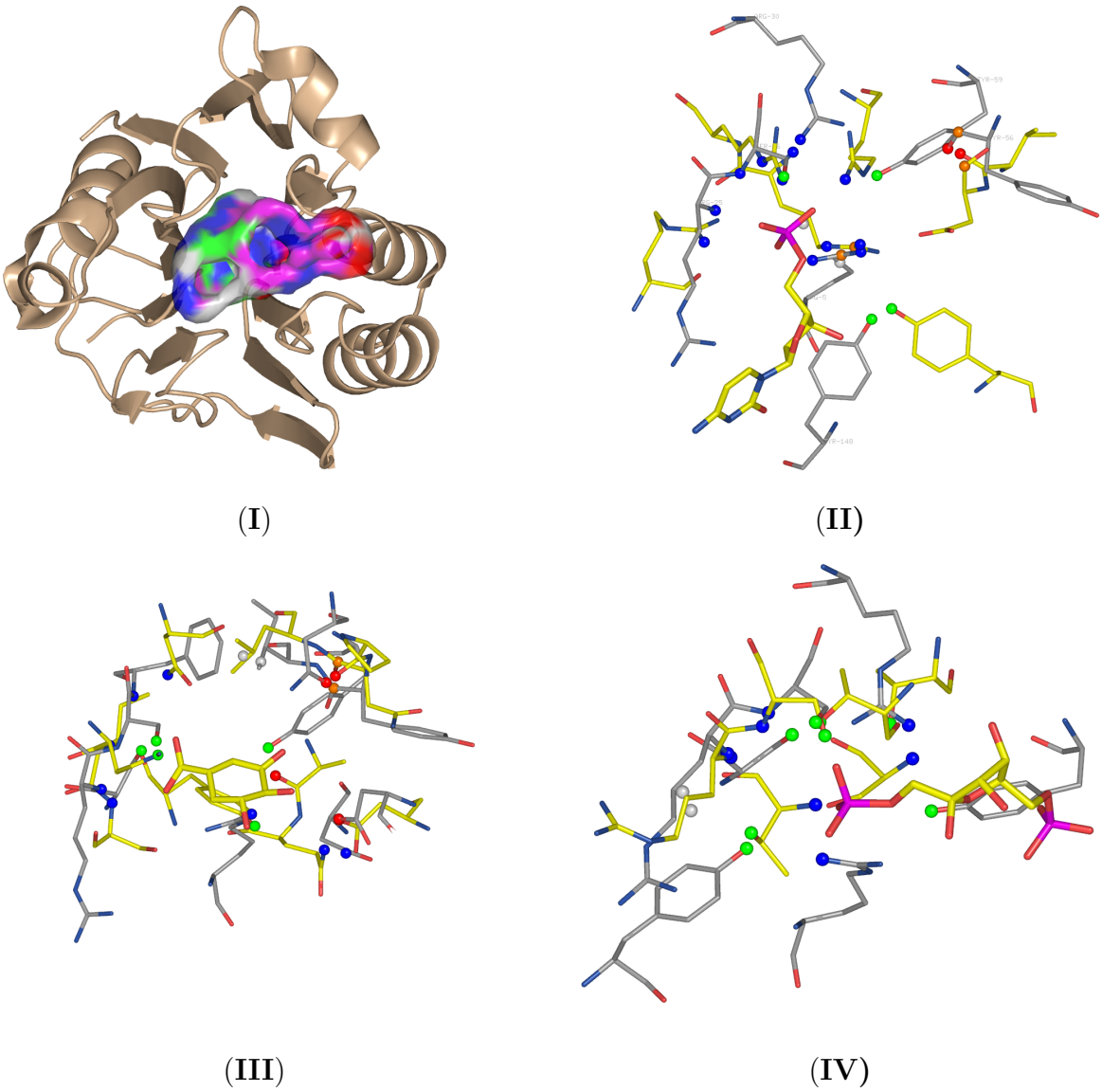


Abb. 4.12 Ein Protein mit unbekannter Funktion, das zur SCOP-Faltungsklasse der Ntn-Hydrolasen gehört. In (I) ist die komplette Bindetasche von Mth1020 (PDB Code 1kuu) dargestellt. (II) zeigt den ersten Rang der Cavbase Ähnlichkeitsanalyse: die Bindetasche einer Deoxycytidylat Hydroxymethylase (PDB Code 1be5, Kohlenstoffe in gelb gefärbt). Die Ähnlichkeiten treten im Bereich der potentiellen Bindetasche von Mth1020 auf. In (III) und (IV) sind Beispiele für Liganden anderer Bindetaschen gezeigt, die als ähnlich gefunden wurden. In (II) ist Typ II Dehydroquinase mit gebundenem Dehydroshikimat dargestellt. In der Cavbase Ähnlichkeitssuche werden viele Bindetaschen mit gebundenem Phosphatgruppen gefunden. In (IV) wird die Überlagerung von Mth1020 mit einer Pyruvat Kinase (PDB Code 1liw, Rang 26) mit einer Phosphatgruppe in der Bindetasche gezeigt. Dies könnte auf eine mögliche Bindungsaffinität von Liganden mit Phosphatgruppen zu Mth1020 hindeuten.

Tab. 4.2 Proteinstrukturen aus *M. thermoautotrophicum*, die 2003 in *structural proteomics* Initiativen aufgeklärt wurden (Daten angepasst nach [Yee et al., 2003]. Die Tabelle enthält Informationen über das zugrundeliegende Gen, die ursprünglich Sequenz-basierte Annotation, die Zuweisung der Proteinfunktion basierend auf Faltungsmustervergleichen, den PDB Code des Proteins, die Anzahl der von Cavbase detektierten Bindetaschen und das Ergebnis der Funktionsannotierung mit Cavbase.

Gen Nummer	ursprüngl. Annotation[1]	struktur. bas. Annotation	PDB	Anz. Taschen	Cavbase Annotierung
Mth0040	RNA polymerase subunit 10	RNA polymerase subunit 10	1EF4	0	keine Tasche detektiert
Mth0129	orotidine decarboxylase	orotidine decarboxylase	1DV7/1DVJ	2	erfolgreiche Annotation
Mth0146	precorrin 8w decarboxylase	precorrin 8w decarboxylase	1F38	10	erfolgreiche Annotation
Mth0150	CHP	NMNATase	1EJ2	1	erfolgreiche Annotation
Mth0152	CHP	FMN-binding protein	1EJE	3	erfolgreiche Annotation
Mth0169	CHP	nucleotide biosynthesis	1GTD	2	keine Annotation
Mth0256	CHP	unbekannt	1NE3	0	keine Tasche detektiert
Mth0538	CHP	response regulatory system	1EIW	1	keine Annotation
Mth0637	CHP	unbekannt	1JRM	0	keine Tasche detektiert
Mth0777	CHP	unbekannt	1KJN	5	keine Annotation
Mth0865	CHP	unbekannt	1IIO	1	keine Annotation
Mth0895	CHP	thioredoxin-like	1ILO	1	keine Annotation
Mth0938	CHP	unbekannt	1IHN	2	mögliche Annotation
Mth1020	CHP	unbekannt	1KUU	1	mögliche Annotation
Mth1048	RNA polymerase H	RNA polymerase H	1EIK	1	keine Annotation

(Fortsetzung nächste Seite)

Fortsetzung Tab. 4.2

Gen Nummer	ursprüngl. Annotation[1]	struktur. bas. Annotation	PDB	Anz. Taschen	Cavbase Annotierung
Mth1175	CHP	unbekannt	1EO1	1	keine Annotation
Mth1184	CHP	unbekannt	1GH9	1	mögliche Annotation
Mth1187	CHP	unbekannt	1LXN	6	keine Annotation
Mth1491	CHP	possible oxido-reductase	1L1S	0	keine Tasche detektiert
Mth1598	CHP	unbekannt	1JW3	1	keine Annotation
Mth1615	CHP	nucleic acid binding	1EIJ	1	mögliche Annotation
Mth1692	CHP	RNA binding	1JCU	2	keine Annotation
Mth1699	CHP	translation elongation factor 1b	1GH8	1	keine Annotation
Mth1743	CHP	ubiquitin-like C-terminal conjugation protein	1JSB R	1	keine Annotation
Mth1747	CHP	dihydroxyacid dehydrogenase	1I36	4	erfolgreiche Annotation
Mth1790	epimerase	epimerase	1EPZ	1	erfolgreiche Annotation

[1] Funktionsannotierung basierend auf einer BLAST Suche gegen eine nicht-redundante Datenbank mit einem E-Wert Cutoff von 10^{-4}; CHP: konserviertes Protein mit unbekannter Funktion (Conserved hypothetical protein).

4.2.2 MJ0577 - ein ATP-bindendes Protein

Eine der ersten Arbeiten auf dem Gebiet der *structural proteomics* war die Aufklärung eines *'hypothetical proteins'* (MJ0577) aus dem thermophilen Bakterium *Methonococcus jannaschii* in der Arbeitsgruppe von Sung-Huo Kim 1998 [Zarembinski et al., 1998]. Die Kristallstruktur wurde mit einem gebundenen ATP-Molekül gelöst, was die Autoren zu der Annahme veranlaßte, daß es sich bei dem unbekannten Protein um eine ATPase handeln könnte. MJ0577 zeigt *in vitro* als isoliertes Protein keine ATPase Aktivität, in Gegenwart von Zelllysat aus *Methonococcus jannaschii* wird ATP aber zu 50 % hydrolisiert. Diese Befunde deuten darauf hin, daß MJ0577 als ATP-vermittelter molekularer Schalter agieren könnte, der ATP in Gegenwart eines weiteren Faktors hydrolisieren kann. MJ0577 zählt aufgrund seiner Sequenz- und Faltungsähnlichkeit zu einer Chaperon-artigen Proteinfamilie; den Universellen Stress Proteinen (*universal stress proteins*, Usp). In der Arbeitsgruppe um David B. McKay konnte die Kristallstruktur eines weiteren Vertreters dieser Familie, das UspA aus *Haemophilus influenzae*, aufgeklärt werden [Sousa and McKay, 2001] (PDB Code 1jmv). Es besitzt eine hohe Faltungsähnlichkeit zu MJ0577 und gehört zur selben SCOP-Superfamilie. UspA ist aber im Gegensatz zu MJ0577 nicht in der Lage ATP zu binden. Beide Proteine stellen zwei Untergruppen der Universellen Stress Protein-Familie dar, verwandt durch Sequenz- und Faltungsähnlichkeit, aber mit unterschiedlichen biochemischen Aktivitäten ausgestattet.

Besitzt MJ0577 strukturelle Ähnlichkeit zu anderen ATP-bindenden Proteinen und kann man aus dem Aufbau der Bindetasche weitere Rückschlüsse auf die Funktion des Proteins erhalten? Um diese Fragestellung zu untersuchen, wurde die ATP-Bindetasche von 1mjh gegen einen Datensatz von 17337 Bindetaschen verglichen (siehe Abbildung 4.13-I). In Tabelle 4.3 sind die ersten Plätze der Vergleichsanalyse gezeigt. Auf den ersten Rängen findet man Strukturen, die ebenfalls ATP gebunden haben und bei denen die ATP-Moleküle beider Strukturen sehr gut überlagern. Die als am ähnlichsten identifizierten Bindetaschen stammen aus Proteinen, die wie MJ0577 zur gleichen SCOP-Superfamilie der Adenin Nukleotid Alpha-Hydrolasen gehören. Zu dieser Superfamilie zählen Proteinfamilien wie die N-Typ ATP Pyrophosphatasen, eine ATPase, eine Phosphoadenylyl Sulfat Reduktase, die α-Untereinheit des Elektronen-Transfer Flavoproteins (ETFP) und die Familie der Universellen Stress Proteine. So wird auf dem ersten Rang die Bindetasche eines ETFP (siehe Abbildung 4.13-II und Tabelle 4.4) gefunden. Die Aminosäuren, die die Bindung zum ATP vermitteln, sind in ETFP

und MJ0577 unterschiedlich. Beide besitzen eine Sequenzidentität von ca 22% und weisen Faltungsähnlichkeiten auf. Auf den weiteren Rängen folgen Proteine, die ebenfalls ATP gebunden haben und zur gleichen SCOP-Superfamilie gehören. Reste, die in die Bindung von ATP verwickelt sind und die ATP-Moleküle überlagern hier ebenfalls gut. Ab Rang 27 werden Proteine gefunden, die keine Faltungsähnlichkeit zu MJ0577 aufweisen. Einige von diesen Proteinen haben ebenfalls ATP oder ähnliche Kofaktoren gebunden. In Abbildung (4.13-III) wird die Überlagerung der ähnlich erkannten Bereiche einer Methyltransferase mit gebundenem S-Adenosyl-L-Homocysteine (PDB Code 1mxi) und MJ0577 gezeigt. Die Adeninbereiche beider Kofaktoren überlagern sehr gut. Abbildung 4.13-IV zeigt die als ähnlich erkannten Bereiche einer Riboflavin Kinase (PDB Code 1n08), die auf dem 56. Rang gefunden wird - beide Proteine zeigen keine Faltungs- und Sequenzhomologie. Es werden aber beim Vergleich mit anderen Proteinen, die ATP oder ATP-artige Liganden gebunden haben, auch Überlagerungen gefunden, in denen Bereiche um die Ribose- und/oder die Phosphatgruppe zur Übereinstimmung gebracht werden. Daraus resultiert eine 'verdrehte' Geometrie der Liganden zueinander ein, wie am Beispiel einer Histidin Kinase (PDB Code 1i5b, siehe Abbildung 4.13-V) gezeigt wird. Neben ATP und ATP-ähnlichen Liganden werden auch eine Reihe von Proteinen als ähnlich gefunden, die Liganden mit einem Molekulargewicht < 500 Dalton gebunden haben. In Abbildung 4.13-VI ist die Überlagerung von MJ0577 mit der Dihydrofolat Reduktase (PDB Code 1diu) komplexiert mit einer Brodimoprim-4,6-Dicarboxylat-Verbindung gezeigt, die freie Aminogruppe am Pyrimidinring überlagert gut mit der Position des Adeninringes bedingt durch ein ähnliches Proteinumfeld.

Cavbase ist in der Lage für Proteine unbekannter Funktion Ähnlichkeiten zu bekannten Proteinstrukturen zu entdecken und so zur Annotation von Proteinstrukturen beizutragen. Darüber hinaus liefert Cavbase wertvolle Ideen über gebundene Liganden oder Ligandfragmente. Dieses Wissen kann Ideen für die Auswahl eines geeigneten Assays zur Funktionsbestimmung geben. Es gibt aber auch Proteinbindetaschen, für die Cavbase keine Ähnlichkeiten findet. In dem untersuchten Datensatz sind keine Bindetaschen ausreichender Ähnlichkeit enthalten. Man kann aber davon ausgehen, daß man mit zunehmender Zahl an aufgeklärten Strukturen in immer mehr Fällen eine erfolgreiche Funktionsannotation aufgrund von Ähnlichkeiten in der Bindetasche, durchführen kann.

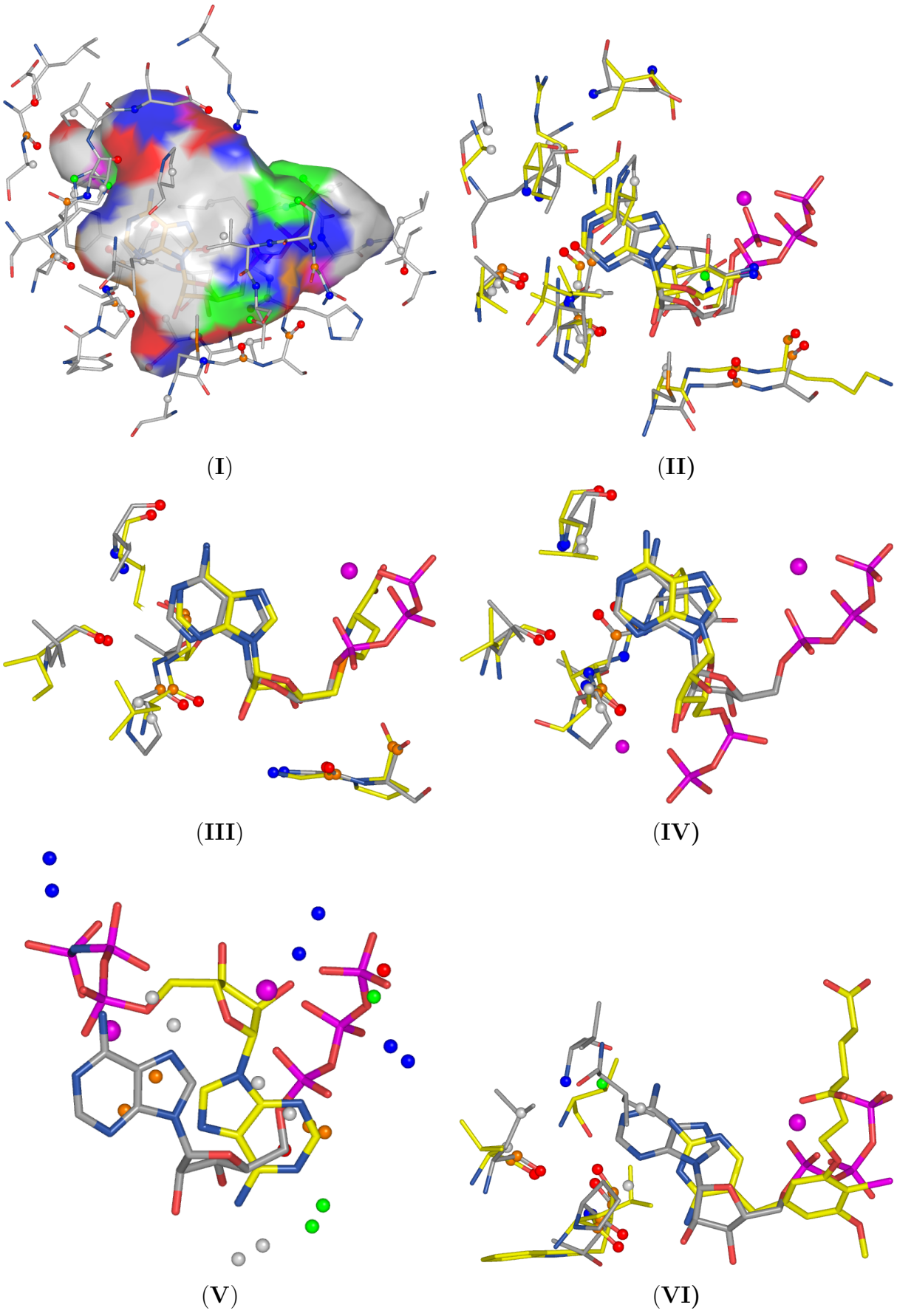


Abb. 4.13 Ergebnisse der Ähnlichkeitsanalyse mit der Bindetaschen von MJ0577 (PDB Code 1mjh). In allen Abbildungen sind die Kohlenstoffe der Aminosäuren und des ATP-Moleküls von MJ0577 in grau gefärbt, die Kohlenstoffe der Proteine sind in gelb dargestellt. (Fortsetzung der Legende siehe nächste Seite.)

Abb. 4.13 (Fortsetzung der Legende) In (I) ist die von Cavbase detektierte Bindetasche von MJ0577 mit gebundenem ATP gezeigt. Das komplette ATP-Molekül wird als zur Bindetasche gehörig erkannt. Abbildung (II) zeigt die Überlagerung von 1mjh mit einem Elektronen-Transfer Flavoprotein (PDB Code 1efp), die in der Ähnlichkeitsanalyse auf dem ersten Rang gefunden wird. (III) zeigt die Überlagerung mit einer Methyltransferase, die S-Adenosyl-L-Homocystein als Kofaktor gebunden hat (PDB Code 1mxi). (IV) stellt die Überlagerung von 1mjh mit einer Riboflavin Kinase (PDB Code 1n08) dar, die zu 1mjh keine Sequenz- und Faltungs-Homologie aufweist. (V) zeigt eine 'verdrehte Geometrie', die aus der Überlagerung ähnlicher Bereiche von 1mjh und einer Histidin Kinase Chea (PDB Code 1i5b) resultiert. In (VI) sind die ähnlichen Bereiche von MJ0577 und der Dihydrofolat Reduktase (PDB Code 1diu) dargestellt, beide Liganden nehmen ähnliche Bereiche in der Tasche ein.

Tab. 4.3 Die ersten 50 Ränge der Ähnlichkeitssuche mit der ATP-Tasche von MJ0577. Werden von einem Protein mehrere Bindetaschen unter den ersten 50 Rängen gefunden, wird jeweils die Tasche mit der besten Bewertung gezeigt.

Rang[1]	Cav ID	Score R1	Ligand[2]	SCOP Score	SCOP-Superfamilie[3]
1	1mjh.1	60.000	ATP-Liganden	50	Adenine nucleotide alpha hydrolases-like
3	1efp.6	19.229	ATP-Liganden	15	Adenine nucleotide alpha hydrolases-like
5	1efv.5	16.777	ATP-Liganden	15	Adenine nucleotide alpha hydrolases-like
6	1o94.13	16.458	ATP-Liganden	15	Adenine nucleotide alpha hydrolases-like
8	1kp3.1	14.432	ATP-Liganden	15	Adenine nucleotide alpha hydrolases-like
9	1kor.6	14.046	ATP-Liganden	15	Adenine nucleotide alpha hydrolases-like
10	1o97.2	13.905	ATP-Liganden	15	Adenine nucleotide alpha hydrolases-like
11	1j1z.1	13.695	ATP-Liganden	15	Adenine nucleotide alpha hydrolases-like
13	1j20.1	13.018	ATP-Liganden	15	Adenine nucleotide alpha hydrolases-like
14	1kh3.1	12.959	ATP-Liganden	15	Adenine nucleotide alpha hydrolases-like
15	1kp2.1	12.906	ATP-Liganden	15	Adenine nucleotide alpha hydrolases-like
18	1j21.1	12.302	ATP-Liganden	15	Adenine nucleotide alpha hydrolases-like
19	1kh2.6	12.227	ATP-Liganden	15	Adenine nucleotide alpha hydrolases-like
27	3prc.7	11.587	unbesetzte ATP-Tasche	0	Bacterial photosystem II reaction centre
31	1k97.1	11.279	unbesetzte ATP-Tasche	15	Adenine nucleotide alpha hydrolases-like
34	1a3w.8	10.744	Ribose/Phosphat	0	Pyruvate kinase, C-terminal domain
38	1cbk.1	10.325	anderer Kofaktor	0	6-hydroxymethyl-7,8-dihydropterin pyrophosphokin.
39	1og1.1	10.312	anderer Kofaktor	-10	No Scop annotation
40	1j3k.8	10.306	anderer Kofaktor	2	Dihydrofolate reductases
41	1iir.1	10.287	unbesetzte ATP-Tasche	2	UDP-Glycosyltransferase-glycogen phosphorylase
42	1asb.3	10.287	kleines Molekül	2	PLP-dependent transferases

(Fortsetzung nächste Seite)

Fortsetzung Tab. 4.3

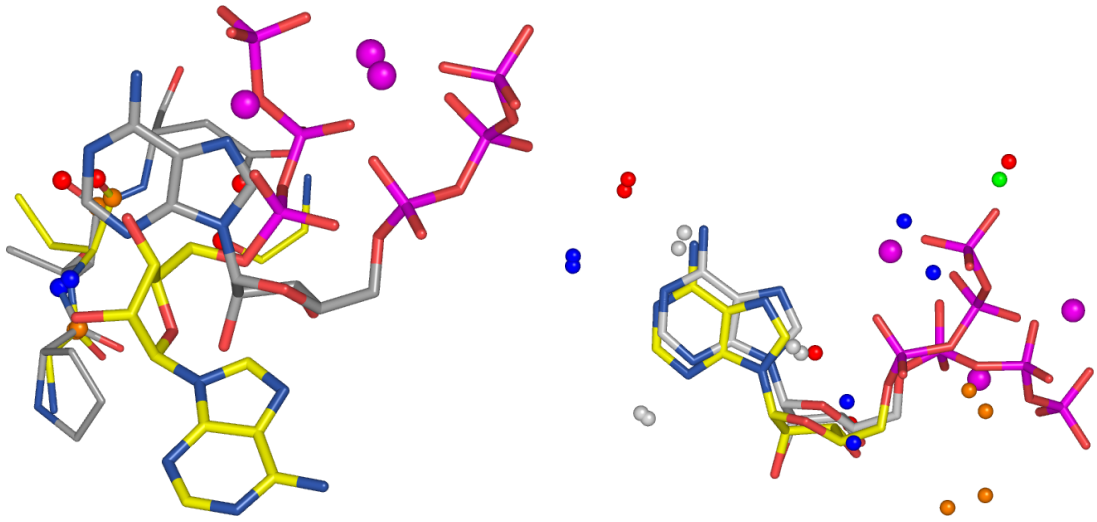
Rang[1]	Cav ID	Score R1	Ligand[2]	SCOP Score	SCOP-Superfamilie[3]
43	1j3j.6	10.254	anderer Kofaktor	2	Dihydrofolate reductases
44	1e0t.8	10.177	unbesetzte ATP-Tasche	0	Pyruvate kinase, C-terminal domain
47	1o01.19	10.075	unbesetzte ATP-Tasche	2	ALDH-like
48	1mbz.12	10.067	ATP-Liganden	15	Adenine nucleotide alpha hydrolases-like
49	1kp5.1	9.904	unbesetzte ATP-Tasche	0	GFP-like
51	1g5q.4	9.74	kleines Molekül	2	DFP DNA/pantothenate metabolism flavoprotein
52	1g5c.15	9.734	kleines Molekül	2	beta-carbonic anhydrase, cab
53	1mxi.1	9.725	anderer Kofaktor	2	alpha/beta knot
54	1o0v.3	9.638	Ribose/Phosphat	0	Immunoglobulin
55	1nhd.6	9.626	anderer Kofaktor	2	No Scop annotation
56	1n08.3	9.625	PIC - ATP-Liganden	0	Riboflavin kinase

[1] Redundante Bindetaschenvergleiche (von mehreren weitgehend identischen Bindetaschen eines Proteins wird nur ein Vergleich gezeigt) sind aus Gründen der Übersichtlichkeit nicht dargestellt.

[2] Die Spalte Ligand beschreibt den gebundenen Liganden und bewertet die Überlagerung des ATP-Molekül aus MJ0577 und dem Liganden des anderen Proteins. Die gefundenen Überlagerungen wurden in folgende Kategorien eingeteilt: **ATP-Ligand** (Adenin/ATP-Liganden beider Taschen überlagern sehr gut); **Ribose/Phosphat** (Ähnlichkeiten werden im Bereich der Ribose/Phosphatbinderegion von MJ0577 gefunden); **kleines Molekül** (ein Wirkstoff-ähnliches Molekül überlagert mit der Region, in der ATP in MJ0577 gebunden wird); **unbesetzte ATP-Tasche** (große Übereinstimmung in der ATP-Tasche, Region wird aber in der zweiten Tasche nicht von einem Liganden besetzt); **anderer Kofaktor** (Bindestelle wird von einem ATP-ähnlichen Kofaktor besetzt, hier mehrere Überlagerungen in der Bindetasche möglich.

[3] In den Fällen, in denen die Bindetasche aus mehreren Proteindomänen aufgebaut ist, die zu unterschiedlichen SCOP-Superfamilien gehören, wird nur eine Domäne angegeben. Angegeben ist die englische Bezeichung der SCOP-Superfamilie.

In der Gruppe von R. Nussinov wurden für MJ0577 mit der ATP-Bindetasche ebenfalls eine Ähnlichkeitssuche durchgeführt und MJ0577 u.a. gegen einen Datensatz mit repräsentativen Proteinstrukturen verglichen [Shulman-Peleg et al., 2004]. Dabei ist die Ähnlichkeit zu der Proteinkinase A (PDB Code 1atp) aufgefallen; die Aminosäuren, die in die Adeninerkennung verwickelt sind, sind in beiden Taschen (Mj0577 und 1atp) zur Übereinstimmung gebracht worden. Diese Ähnlichkeit läßt sich mit Cavbase unter Verwendung von Standard-Parametern nicht einfach reproduzieren. Die beiden Taschen werden so überlagert, daß die Ähnlichkeit im Bereich der Phosphatgruppen beider ATP-Moleküle gefunden wird. Die ATP-Moleküle sind um 180 Grad 'gedreht' (siehe Abbildung 4.14-I). Erweitert man den Suchraum (betrachtete Clique-Lösungen) und erlaubt, daß Oberflächenbereiche als ähnlich angesehen werden, wenn sie zu mindestens 60 % überlappen (overlap_surfpatch, Standardwert: mindestens 70 %) wird eine Lösung gefunden, in der beide ATP-Moleküle sehr gut überlagert werden, in der die Bereiche in der Bindetasche als ähnlich erkannt werden (siehe Abbildung 4.14-II).



(**I**) (**II**)

Abb. 4.14 Einfluß der Parameterwahl bei der Ähnlichkeitsanalyse auf die erhaltene Überlagerung der Bindetaschen von MJ0577 und einer Proteinkinase (PDB Code 1atp). Verwendet man zum Vergleich bei der Ähnlichkeitssuche von 1mjh und 1atp, erhält man kein signifikantes Ähnlichkeitsmaß zwischen beiden Taschen, das auf eine Verwandtschaft zwischen beiden Taschen schließen lässt. Die gebundenen Liganden werden unbefriedigend überlagert, die als ähnlich angesehenen Bereiche erstrecken sich im Bereich der Phosphatbinderegion. Verwendet man für die Vergleichsanalyse Parameter, die einen großen Suchraum abdecken, wird eine sehr gute Überlagerung gefunden, die der erhaltenen Überlagerung von Nussinov et. al ähnlich ist (II).

Tab. 4.4 Als äquivalent erkannte Pseudozentren und Aminosäuren von MJ0577 und dem Elektronen-Transfer Flavoprotein.

MJ0577 (1mjh.1)			**ETFP**(1efp.6)		
Typ des Pseudo-zentrums	äquivalente Aminosäure[a]		Typ des Pseudo-zentrums	äquivalente Aminosäure[a]	
Pi	P 1011	p	Pi	P 6	p
Akzeptor	P1011	p	Akzeptor	P 6	p
Pi	T 1012	p	Pi	V 7	p
Donor	T 1012	p	Donor	V 7	p
Akzeptor	T 1012	p	Akzeptor	V 7	p
Donor-Akzeptor	S 1015	s	Donor	N 36	s
Pi	L 1039	p	Pi	V 61	p
Akzeptor	L 1039	p	Akzeptor	V 61	p
Donor	V 1041	p	Donor	I 63	p
Donor	D 1043	p	Donor	I 96	p
Pi	G 1127	p	Pi	G 120	p
Akzeptor	G 1127	p	Akzeptor	G 120	p
Pi	S 1128	p	Pi	K 121	p
Akzeptor	S 1128	p	Akzeptor	K 121	p
Donor	V 1142	p	Donor	T 131	p
Aliphatisch	P 1011	s	Aliphatisch	P 6	s
Aliphatisch	L 1039	s	Aliphatisch	V 61	s
Aliphatisch	A 1081	s	Aliphatisch	A 68	s
Aliphatisch	K 1084	s	Aliphatisch	R 9	s
Aliphatisch	P 1108	s	Aliphatisch	V 101	s
Aliphatisch	I 1112	s	Aliphatisch	L 105	s
Aliphatisch	M 1126	s	Aliphatisch	A 119	s

[a] Ein-Buchstabencode der Aminosäure, Aminosäurenummer und Ursprung des Pseudozentrums: aus der Seitenkette (s) oder aus der Peptidbindung (p).

4.3 Vorhersage der Kreuzreaktivität von Celecoxib an Carboanhydrase

4.3.1 Inhibitoren der Carboanhydrase zeigen ebenfalls Bindungseigenschaften an Cyclooxygenasen

Das unterschiedliche klinische Nebenwirkungsprofil der Cyclooxygenase-Inhibitoren Celecoxib (**2**) und Rofecoxib (**4**) [FitzGerald and Patrono, 2001] legte nahe, dass neben der Cyclooxygenase-2 (COX-2) Inhibition als Hauptwirkung unterschiedliche pharmakologische Wirkungen dieser Verbindungen auftreten. So zeigt Celecoxib - nicht aber Rofecoxib - u.a. eine anti-ödemative und blutdrucksenkende Wirkung [Whelton, 2001; Whelton et al., 2002]. Beide Verbindungen sind sich strukturell sehr ähnlich, Celecoxib besitzt aber eine terminale Sulfonamid-Gruppe, während Rofecoxib hier eine Methylsulfongruppe aufweist (siehe Abbildung 4.15). Klassische Inhibitoren der Carboanhydrasen (CA) besitzen ebenfalls eine freie terminale Sulfonamid-Gruppe, mit der sie eine starke Bindung an das katalytische Zink vermitteln. Die unterschiedlichen pharmakologischen Eigenschaften der beiden COX-2-Hemmer könnten auf einer Hemmung der CAs durch Celecoxib beruhen. Um diese Vermutung zu überprüfen, wurden in einer Kooperation mit der Arbeitsgruppe von Prof. Dr. C. Supuran (Universität Florenz) die Affinitäten klassischer CA- und COX-2-Inhibitoren an CA getestet [Weber et al., 2004]. Es wurden Inhibitionsdaten von COX-2-Inhibitoren bezogen auf vier CA Isoenzyme bestimmt. Dabei zeigt sich, daß COX-2-Inhibitoren, die keine freie Sulfonamidgruppe besitzen (wie SC125 (**5**), Rofecoxib (**4**), Diclofenac), die CA nicht hemmen, während Celecoxib ((**2**) hochaffin an CA-II, CA-III und CA-IV bindet (siehe [Weber et al., 2004], Tabelle 1). Diese Befunde konnten ebenfalls im Tiermodell (Senkung des Augeninnendrucks am Kaninchenauge) bestätigt werden.

Die Kristallstruktur von Celecoxib wurde im Komplex mit Carboanyhdrase-II in unserer Arbeitsgruppe aufgeklärt (PDB Code 1oq5) [Weber et al., 2004]. Der Bindungsmodus ähnelt dem der klassischen Carboanhydraseinhibitoren, die Sulfonamidgruppe koordiniert an das Zinkatom, weitere wichtige Wasserstoffbrücken zwischen Inhibitor und Enzym werden entsprechend den bekannten Inhibitoren ausgebildet.

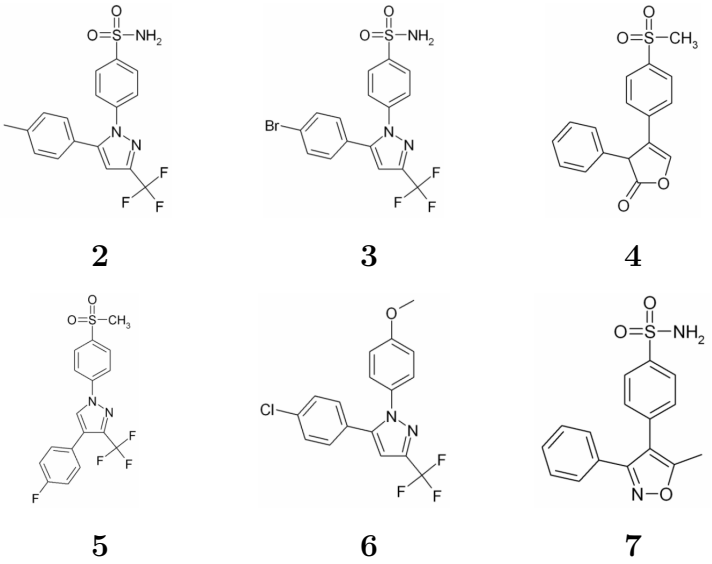


Abb. 4.15 Inhibitoren der Cyclooxygenase: Celecoxib (2), SC558 (3), Rofecoxib (4), SC125 (5), SC560 (6)und Valdecoxib (7).

4.3.2 Struktureller Vergleich von CA-II mit COX-2

Carboanhydrasen und Cyclooxgenasen zeigen keine Sequenz-, Faltungs- oder funktionelle Ähnlichkeit. In Abbildung 4.16 ist das Faltungsmuster beider Proteine gezeigt, die Carboanhydrase besitzt als Grundstruktur ein zehnsträngiges β-Faltblatt, während die Cyclooxygenase eine überwiegend α-helikale Faltung zeigt. Die Bindetasche

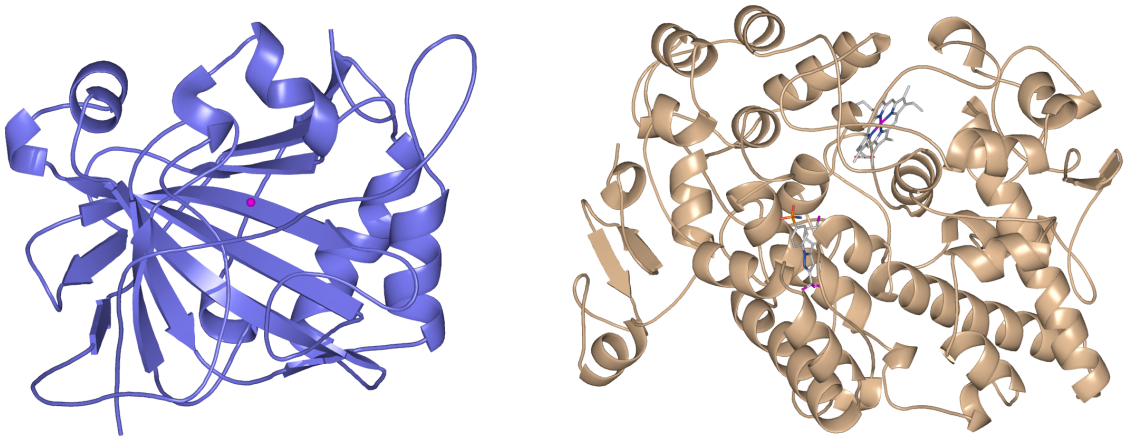


Abb. 4.16 Faltungsmuster der Carboanhydrase II und des Cyclooygenase-2 Monomers. In (I) ist das zehnsträngige β-Faltblatt der Carboanhydrase gezeigt, in (II) die überwiegend α-helikale Faltung eines COX-2 Monomers dargestellt.

von Celecoxib in COX-2 ist sehr groß, da sie sich an der Schnittstelle zwischen zwei COX-2-Monomeren befindet und neben der COX-Bindestelle auch noch die Peroxidasebindestelle umfasst. Aus diesem Grund wurden für den strukturellen Vergleich zuerst Subtaschen verwendet, die nur die COX-Bindestelle umfassen. Die Anfragebindetasche hat einen zu Celecoxib analogen Inhibitor, SC-558 (**3**), gebunden. SC-558 besitzt anstelle eines p-Bromphenyl-Substituenten einen p-Methylphenyl-Substituenten. Bioisosterie der beiden Gruppen kann in diesem Fall als gegeben angesehen werden. Celecoxib besitzt ein rigides Grundgerüst. Man erhält eine gute Überlagerung der Inhibitoren aus beiden Kristallstrukturen basierend auf den Atomen der drei Ringsysteme (siehe Abbildung 4.17-IV). Die Orientierung der Sulfonamidankergruppe unterscheidet sich aber in beiden Strukturen, was einen Vergleich der kompletten Bindetaschen erschwert. Die exponierten Eigenschaften der Donor- (NH) und Akzeptor- (O) Gruppen der Sulfonamidgruppe sind im Bezug zu der Ebene, die durch den Schwefel der Sulfonamidgruppe und dem benachbarten Phenylring aufgespannt wird, gespiegelt. Aus diesem Grund wurden die drei Subtaschen, die den Sulfonamidanker, die Trifluormethylgruppe und den p-Bromphenylring von Celecoxib beherbergen, separat gegen einen Datensatz von 9433 Bindetaschen verglichen (siehe Abbildung 4.17-I bis III). Benutzt man die Sulfonamidsubtasche (bestehend aus 25 Pseudozentren) als Anfragetasche, findet man auf Rang 38 die erste Tasche aus einer Carboanhydrase (d.h. nachdem man 0.4% des Datensatzes durchmustert hat). Weitere Carboanhydrasen werden auf den folgenden Rängen gefunden. Die Unterschiede in den Scoringwerten zwischen den gefundenen Carboanhydrasen sind unter Berücksichtigung beider Scoringfunktionen R_1 und R_2 gering, obwohl sie teilweise große Unterschiede in den Rängen aufweisen. Dieser Umstand erschwert ein sinnvolles abgestuftes Sortieren der erhaltenen Vergleiche, da kleine Änderungen in den Scoringwerten einen großen Einfluß auf das Vergleichsranking besitzen. Überlagert man aber die gefundenen paarweisen Vergleiche, wird die Sulfonamidgruppe von SC-558 mit den Sulfonamidgruppen der CA-Inhibitoren aus verschiedenen Bindetaschen gut überlagert (siehe Abbildung 4.17-II).

Die Bindetasche, die den Trifluormethylrest in der COX-2 Struktur zeigt, wird durch sieben Pseudozentren beschrieben (siehe Abbildung 4.17-III) und hat einen vorwiegend aliphatischen Charakter. Vergleicht man diese Subtaschen gegen den Datensatz aus den 9433 Bindetaschen, findet man unter den ersten 200 Rängen 41 Carboanhydrasen, die Liganden aufweisen, die in einen ähnlichen Bereich wie die Trifluormethylgruppe von SC-558 binden. Interessanterweise wird auf Rang 76 (d.h. nachdem man 0.8% des Datensatzes durchmustert hat) eine CA II Struktur mit einem gebundenen Sulfonamid-

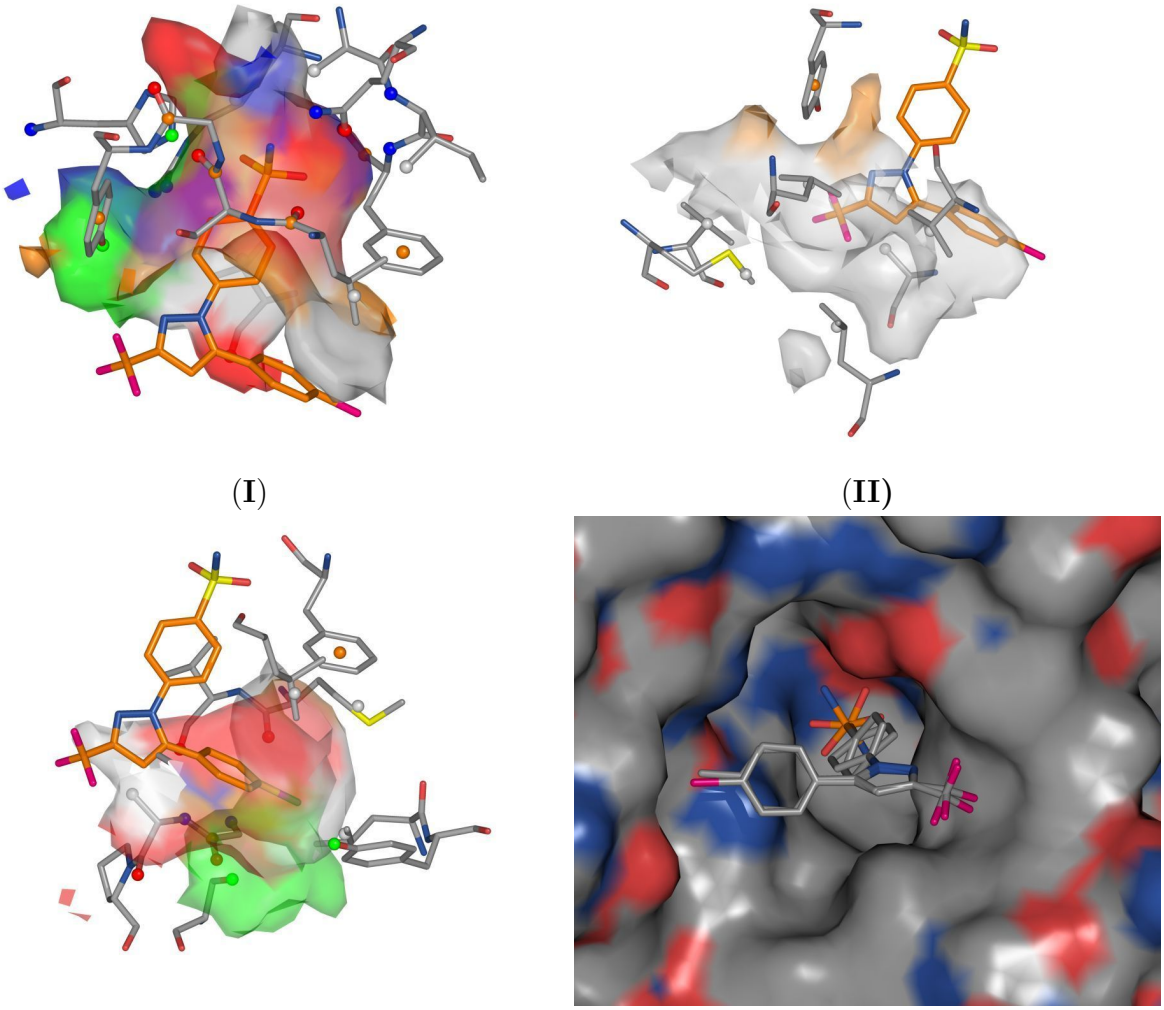
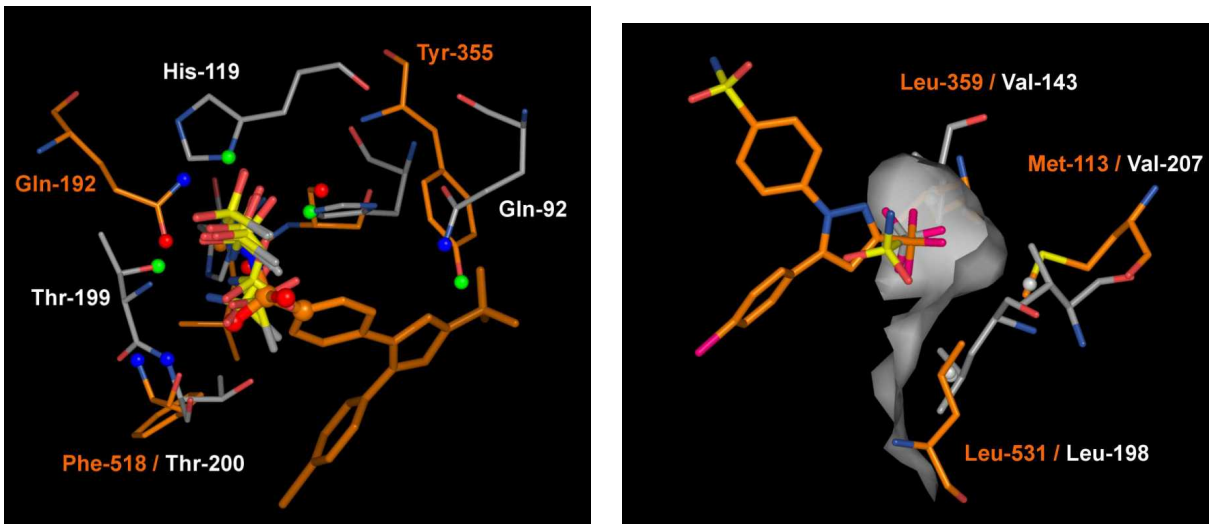


Abb. 4.17 Subtaschen der COX-2 Struktur im Komplex mit Celecoxib-Analogon. In (I) sind die Aminosäuren, Pseudozentren und die Bindetaschenoberfläche der Sulfonamid-Subtasche gezeigt, in (II) die Subtasche um die Trifluormethylgruppe, in (III) die Subtasche, die die 4-Methyl-phenyl beherbergt. Mit diesen Subtaschen wurden die Vergleichsanalysen durchgeführt. In (IV) ist die Überlagerung von SC-558 aus COX-2 mit Celecoxib aus Carboanhydrase gezeigt. Die Grundgerüste überlagern gut, man erkennt aber ein verdrehtes Arrangement der Sulfonamidgruppe, was eine Ähnlichkeitsanalyse mit der gesamten Tasche erschwert [Weber et al., 2004]

basierten Inhibitor gefunden, der ebenfalls eine Trifluormethylgruppe als Rest besitzt. Beide Trifluormethylgruppen werden sehr gut überlagert, d.h. die Liganden orientieren diese Gruppe in vergleichbarer Weise in diese Subtasche (siehe Abbildung 4.18-II). Tabelle 4.6 zeigt die als ähnlich gefundenen Pseudozentren und zugehörigen Aminosäuren. Die Bindetasche, die den p-Bromphenylrest trägt, wird durch 14 Pseudozentren charakterisiert. Für diese Subtasche wurde auf den ersten Rängen keine CA Bindetasche



(**I**) (**II**)

Abb. 4.18 Vom Vergleichsalgorithmus als ähnlich angesehene Bereiche in den Bindetaschen von CA und COX. In (I) ist eine Überlagerung des COX-2 Inhibitors SC-558 mit den Sulfonamidgruppen verschiedener gefundener Inhibitoren der CA Isoenzyme gezeigt. Zusätzlich dargestellt sind die Aminosäuren und die Pseudozentren. In (II) ist die Überlagerung der Trifluormethyl-Subtasche mit der CA-II Struktur (PDB Code 1bcd) gezeigt. Die Trifluormethylgruppen beider Inhibitoren überlagern sehr gut.

gefunden. Dieses Ergebnis ist nicht überraschend, wenn man den Charakter und die Umgebung beider Taschen vergleicht. Die Subtasche in COX-2 ist tief im Protein vergraben und besteht aus aromatischen Aminosäuren (Phe-518, Trp-387, Tyr-385, Phe-381), während die Tasche in CA II teilweise Solvens-exponiert ist und aus hydrophilen Aminosäuren (Asn-67, Glu-69, Gln-92) zusammengesetzt ist. Beide Bindetaschen weisen ziemlich unterschiedliche physikochemische Eigenschaften auf, eine Ähnlichkeit ist mit unserem Ansatz nicht zu entdecken.

Die kombinierte Auswertung der drei separaten Subtaschen macht deutlich, daß Cavbase in der Lage ist, ähnliche Bereiche in den Bindetaschen von CA-II und COX-2 zu identifizieren. In zwei der drei Subtaschen wird ein physikochemisch vergleichbares Umfeld entdeckt. Die als ähnlich erkannten Bereiche sind in beiden Taschen aber relativ klein, so daß sie sich im Fall der Sulfonamidsubtasche nicht ausreichend vom Rest des Datensatzes absetzen können. Kombiniert man allerdings die Scoringwerte für die Trifluormethyl- und die Sulfonamidtasche, indem man beide Werte addiert und den Datensatz neu sortiert, findet man die erste Carboanhydrase-Tasche auf Rang 3 (PDB Code 1bn4). Tabelle 4.8 zeigt die besten 50 Ränge nach dem kombinierten Bewertungsschema. Auf den ersten 11 Plätzen sind erwartungsgemäß COX-2 verwandte Proteine

Tab. 4.5 Als äquivalent erkannte Pseudozentren und Aminosäuren der Sulfonamid-Bindetaschen von COX-2 und CA II

Cyclooxygenase (6cox.97)			**Carboanhydrase** (1bn4.1)		
Typ des Pseudo-zentrums	äquivalente Aminosäure[a]		Typ des Pseudo-zentrums	äquivalente Aminosäure[a]	
Akzeptor	Q192	s	Donor-Akzeptor	T199	s
Donor	Q192	s	Donor-Akzeptor	H119	s
Pi	L352	p	Aromatisch	H96	s
Akzeptor	L352	p	Donor-Akzeptor	H96	s
Akzeptor	S353	p	Donor-Akzeptor	H94	s
Donor-Akzeptor	Y355	s	Donor	Q92	s
Donor	F518	p	Donor	T200	p

[a] Ein-Buchstabencode der Aminosäure, Positionsnummer und Lokalisierung des Pseudozentrums: auf der Seitenkette (s) oder auf der Peptidbindung (p).

Tab. 4.6 Ähnlichkeitssuche mit der Trifluormethyl-Bindetasche. Als äquivalent erkannte Pseudozentren und Aminosäuren von COX-2 und CA II.

Cyclooxygenase (6cox.93)			**Carboanhydrase** (1bcd.1)		
Typ des Pseudo-zentrums	äquivalente Aminosäure[a]		Typ des Pseudo-zentrums	äquivalente Aminosäure[a]	
Aliphatisch	M113	s	Aliphatisch	V207	s
Aliphatisch	L359	s	Aliphatisch	V143	s
Aliphatisch	L531	s	Aliphatisch	L198	s

[a] Ein-Buchstabencode der Aminosäure, Positionsnummer und Lokalisierung des Pseudozentrums: auf der Seitenkette (s) oder auf der Peptidbindung (p).

zu finden. Auffällig oft werden auf den nächsten Rängen Chaperon-Proteine wie das *Heat Shock Protein 70* gefunden. Das Heat Shock Protein 70 existiert in der konstitutiven (Hsc70) und in der induzierbaren Isoform (Hsp70) [Doong et al., 2002], beide sind sich sehr ähnlich. In Abbildung 4.19 ist eine Überlagerung der Cyclooxygenase mit dem Hsc70 (PDB Code 1kaz) dargestellt und Tabelle 4.7 zeigt die Pseudozentren und Aminosäuren, die in räumliche Übereinstimmung gebracht wurden. Beide Bindetaschen zeigen Ähnlichkeiten im Bereich der Sulfonamidbindestelle der COX-2. Interessanterweise werden die Ähnlichkeiten genau in dem Bereich von Hsc70 entdeckt, der bei der Protein-Protein-Interaktion von Hsc70 bzw. Hsp70 mit dem Bag-1 Protein eine wichtige Rolle spielt [Sondermann et al., 2001]. Das Bag-1 (Bcl2- associated athanogene 1) Protein ist ein Kofaktor der Chaperone der Hsp70 Familie. Wenn Bag-1 an die Hsp70 ATPase Domäne bindet, kommt es zu einer ATP-abhängigen Freisetzung des Substrates von Hsp70. Die Hsp70/Hsc70 ATPase- und die Bag-1-Domäne bilden einen Komplex, an dessen Schnittstelle den Aminosäuren Arg261 und Glu283 von Hsc70 eine besondere Bedeutung zukommt (siehe Abbildung 4.19-I). Beide Aminosäuren sind in allen cytosolischen und eukaryotischen Hsp70 Proteinen zu finden. Cavbase findet hier in diesem Bereich eine Ähnlichkeit von Hsc70 mit der Sulfonamidtasche von COX-2. Möglicherweise ist Celecoxib in der Lage, in diesem Bereich zu binden und auf die Interaktion von Hsc70 und Bag-1 einzuwirken.

Tab. 4.7 Ähnliche Reste in den Bindetaschen von Hsc70 (PDB Code 1kaz) und COX-2(PDB Code 6cox).

Hsc70 (1kaz.1)			**COX-2** (6cox.97)		
Typ des Pseudozentrums	äquivalente Aminosäure[a]		Typ des Pseudozentrums	äquivalente Aminosäure[a]	
Pi	K56	p	Pi	L352	p
Akzeptor	K56	p	Akzeptor	L352	p
Pi	N57	p	Pi	S353	p
Akzeptor	N57	p	Akzeptor	S353	p
Akzeptor	R261	p	Akzeptor	F518	p
Aliphatisch	K56	s	Aliphatisch	L352	s
Aliphatisch	R262	s	Aliphatisch	R513	s
Aliphatisch	T265	s	Aliphatisch	V523	s

[a] Ein-Buchstabencode der Aminosäure, Positionsnummer und Lokalisierung des Pseudozentrums: auf der Seitenkette (s) oder auf der Peptidbindung (p).

Tab. 4.8 Kombinierte Scoringwerte für die Sulfonamid-Subtasche und die Trifluormethylgruppe. Die ersten 50 Ränge der kombinierten Scoringwerte für die Sulfonamidsubtasche und die Trifluormethylsubtasche.

Rang	Cav ID	Score CF3	Score Sulfon	EC-Nummer[1]
1	6cox.2	7	21	1.14.99.1
2	1ddx.1	6.58	17.48	1.14.99.1
3	1cvu.2	5.11	18.32	1.14.99.1
4	1cx2.4	6.08	15.54	1.14.99.1
5	1cx2.1	6.08	14.53	1.14.99.1
6	1ht8.2	6.18	12.94	1.14.99.1
7	5cox.1	4.28	14.7	1.14.99.1
8	1igz.1	5.32	10.59	1.4.99.1
9	1diy.5	4.2	11.39	1.14.99.1
10	1fe2.1	3.41	8.86	1.14.99.1
11	1hpm.1	0.78	7.58	3.6.1.3
12	1c7o.11	1.49	6.25	—
13	1bn4.1	2.28	5.46	4.2.1.1
14	1phk.1	1.54	5.94	2.7.1.38
15	1bup.1	0.83	6.59	—
16	1kaz.1	0.8	6.62	—
17	1ba0.1	0.81	6.55	—
18	5cel.1	1.44	5.88	3.2.1.91
19	1ci3.1	2.57	4.75	—
20	1gek.2	1.65	5.63	1.14.15.1
21	1iem.1	2.3	4.97	—
22	1bxr.6	2.42	4.84	6.3.5.5
23	2hhe.4	3.15	4.09	—
24	1ngb.1	1.48	5.67	—
25	1d3g.2	2.4	4.74	—
26	1c7o.3	0.85	6.28	—
27	1ffv.2	1.71	5.4	—
28	1c8j.2	2.32	4.76	—
29	2rm2.12	1.53	5.54	3.4.22.28
30	1k0i.2	1.5	5.57	—
31	1a0v.5	2.45	4.6	—
32	1c7d.3	2.98	4.06	—

(Fortsetzung nächste Seite)

Fortsetzung Tab. 4.8

Rang	Cav ID	Score CF3	Score Sulfon	EC-Nummer[1]
33	1jg2.1	1.51	5.52	—
34	1gnj.5	2.27	4.75	—
35	1hjo.1	0.88	6.13	—
36	1fr6.2	0.81	6.19	—
37	1kay.1	0.8	6.16	—
38	1rkd.1	1.9	5.06	2.7.1.15
39	1kyq.5	0.82	6.13	—
40	1at1.8	1.45	5.5	2.1.3.2
41	1ewh.1	2.55	4.39	—
42	1htt.7	0.79	6.14	6.1.1.21
43	1fft.28	1.75	5.18	—
44	1qtn.1	0.84	6.08	—
45	1dke.2	3.06	3.82	—
46	1qdu.18	1.73	5.14	—
47	1czj.1	1.47	5.39	—
48	1daj.1	2.38	4.43	1.5.1.3
49	1hgc.4	2.5	4.31	—
50	1fpc.1	0.96	5.85	3.4.21.5

[1] Die Zuordnung Proteinstruktur zu Enzymklassifikation wurde der Protein Datenbank (PDB) entnommen. Fehlt die Annotierung oder handelt es sich bei dem Protein um kein Enzym, kann keine Angabe erfolgen (-).

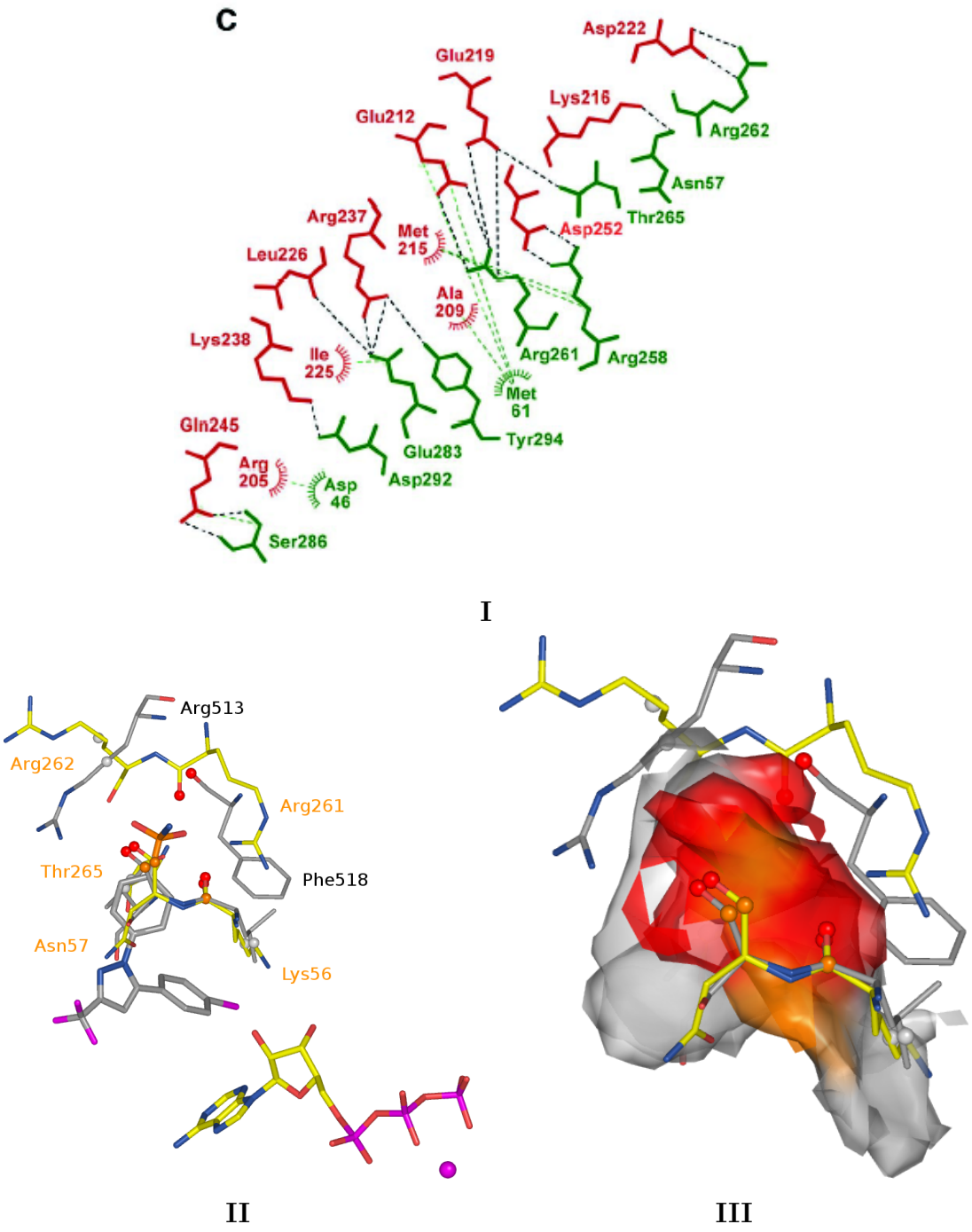


Abb. 4.19 Aminosäuren, die bei der Protein-Protein-Interaktion von Hsc70 mit dem Bag-1 Protein (PDB Code 1hx1) eine wichtige Rolle spielen und ähnliche Bereiche in den Bindetaschen von COX-2 (PDB Code 6cox, Sulfonamidsubtasche) und dem Hsc70 (PDB Code 1kaz). In (I) sind die Aminosäuren gezeigt, die im Bag-1 und Hsc70 Protein-Protein-Komplex den Kontakt ausbilden (Abbildung entnommen aus [Sondermann et al., 2001]). In rot sind die Aminosäuren des Bag-1 Proteins, in grün die des Hsc70 gezeigt, gestrichelte Linien geben Wasserstoffbrücken wieder, schraffierte Halbkreise deuten hydrophobe Wechselwirkungen an. (II) zeigt die als ähnlich erkannten Bereiche in den Bindetaschen von COX (Kohlenstoffe in grau gefärbt, Beschriftung in schwarz) und Hsc70 (Kohlenstoffe gelb gefärbt, Beschriftung in orange), in (III) sind die Aminosäuren, Pseudozentren und gemeinsame Oberflächenbereiche aus beiden Bindetaschen dargestellt. Die Ähnlichkeit wird in dem Bereich von Hsc70 detektiert, in dem die Aminosäuren von Bag-1 und Hsc70 interagieren.

4.4 Malatdehydrogenase (MDH) und Carboanhydrase (CA) besitzen ähnliche Bereiche in der Bindetasche

4.4.1 Inhibitoren der Carboanhydrase (CA) zeigen Aktivität an Malatdehydrogenasen (MDH)

Der Arzneistoffklasse der Sulfonamide wird eine Vielzahl von pharmakologisch nützlich Effekten nachgewiesen, so daß sie in verschiedenen Indikationsgebieten (als Diuretika, Antibiotika, etc.) in der klinischen Anwendung erfolgreich eingesetzt werden. Kürzlich wurde das Antitumor-Potential neuartiger Sulfonamidderivate entdeckt und konnte durch *in vitro* und *in vivo* Studien belegt werden [Casini et al., 2002]. Beispielsweise zeigen die Disulfonamidphenyl-Verbindung E7070 und abgeleitete Derivate eine gute Inhibition verschiedener Krebszelllinien [Owa et al., 1999; Ozawa et al., 2001; Supuran, 2003]. Der Wirkungsmechanismus ist multifaktoriell und noch nicht vollständig aufgeklärt. E7070 agiert in der G1 Phase des Zellzyklus und verhindert die Phosphorylierung von Cyclin E und die Aktivierung der CDK-2 [Dittrich et al., 2003; Van Kesteren et al., 2002]. Dadurch wird ein Fortschreiten des Zellzyklus in die S-Phase verhindert. Zur Zeit befindet sich die Verbindung E7070 (**8**) in klinischen Studien (Phase II) zur Therapie einiger Tumorformen. Für Inhibitoren mit einer freien terminalen Sulfonamidgruppe ist eine potente Inhibition von Carboanhydrasen zu erwarten. Supuran und Mitarbeiter konnten eine nanomolare Inhibition von E7070 an Carboanhydrasen experimentell belegen und außerdem den Bindungsmodus von E7070 aufklären [Abbate et al., 2004]. E7070 bindet in ähnlicher Weise wie bekannte Carboanhydraseinhibitoren mit der terminalen Sulfonamidgruppe an das katalytische Zinkion im aktiven Zentrum der Carboanhydrase.

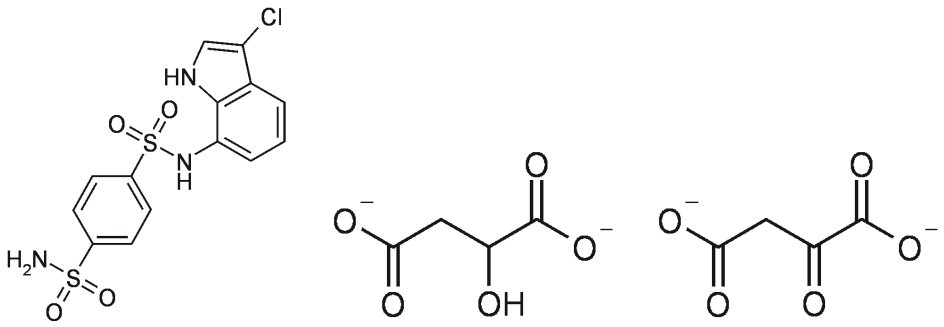
Durch die experimentelle Analyse des Proteoms von Zellen zu verschiedenen Zeiten sind Oda et al. in der Lage, Proteine aus dem Zelllysat herauszufischen, die Bindung an einen vorgegebenen Inhibitor zeigen. Dazu wird dieser Inhibitor auf einem festen Träger immobilisiert. In einem solchen Experiment wurde E7070 (und Derivate) über einen Linker auf einem polymeren Träger verankert. Die freie Sulfonamidgruppe von E7070 wurde durch eine Methylaminofunktion ersetzt, um an den Linker gebunden zu werden. Verschiedene Zelllysate wurde über die präparierten Matrizen eluiert

und gebundene Proteine mittels 2D-Protein Gelelektrophorese, Massenspektrometrie und HPLC (Hoch-Druck-Flssigkeits-Chromatographie) untersucht. Zusätzlich wurde die zentrale Sulfonamidgruppe, die den Indolring mit dem Phenylring verbindet, durch eine Amidgruppe ersetzt und am Linker fixiert. Solche Inhibitoren zeigen keine große Aktivität gegenüber den verwendeten Zelllinien und sollen bei der Unterscheidung von spezifischer und unspezifischer Aktivität unterstützen. So konnten Proteine identifiziert werden, die nicht nur unspezifisch mit E7070 und dessen Derivaten interagieren. Dabei ist die cytosolische Malatdehydrogenase (cMDH) als spezifisch bindend aufgefallen. Außerdem wurde der Einfluß von E7070 auf die Transkription in bestimmte Krebszellen in DNA-Microarray-Experimenten untersucht und eine direkte Interaktion von E7070 mit cMDH durch Oberflächen-Plasmon-Resonanz-Spektroskopie (SPR-Spektroskopie) bestätigt. Beide Experimente deuten darauf hin, daß cMDH ein *Target* von E7070 ist, allerdings zeigt E7070 in einem Standard-Affinitätsassay keine Aktivität an cMDH. Die Autoren führen diese Beobachtung auf die schlechte Löslichkeit von E7070 zurück [Oda et al., 2003]. Gibt es aber in den Bindetaschen dieser beiden Proteine Gemeinsamkeiten, die den Verdacht auf ein ähnliches Bindungsverhalten an beiden Proteinen aus strukturellen Gründen erhärten könnten? Um diese Hypothese zu überprüfen, wurden die Bindetaschen von Carboanhydrasen und Malatdehydrogenasen miteinander verglichen.

4.4.2 Struktur und Funktion von MDH

Die Malatdehydrogenase ist ein Enyzm, das in den Citratzyklus involviert ist und die Umwandlung von Malat (**9**) zu Oxaloacetat (**10**) katalysiert. Es spielt außerdem eine wichtige Rolle in der reduktiven Carboxylierung von Pyruvat zu Oxalacetat (siehe Abbildung 4.20).

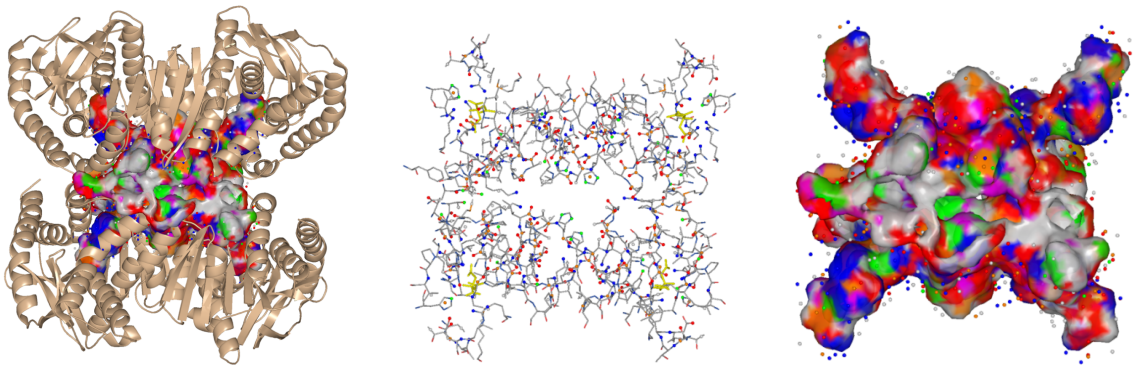
Enzyme des Citratzyklus sind in der inneren Membran der Mitochondrien lokalisiert (mitochondrial malate dehydrogenase, [mMDH]). In Eukaryonten sind noch weitere Formen der Malatdehydrogenase bekannt, die in anderen zellulären Kompartimenten anzutreffen sind (cytosolic malate dehydrogenase, [cMDH]). Trotz einer geringen Sequenzidentität zwischen cMDH und mMDH ist die strukturelle Ähnlichkeit sehr groß, beide Enzyme weisen eine Domäne mit Rossman-Faltung auf. MDHs kristallisieren mit zwei oder vier Monomeren in der asymmetrischen Zelle und liegen vermutlich in Lösung als Dimere vor [Gleason et al., 1994; Chapman et al., 1999]. Prokaryotische Zellen, denen die Zellkompartimentierung fehlt, besitzen nur eine Form der MDH. MDH ge-



8 **9** **10**

Abb. 4.20 Chemische Formeln von E7070 (Indisulam) (8) und natürlichen Substraten der MDH: Malat (9) und Oxaolacetat (10).

hört zur Gruppe der NAD-abhängigen Dehydrogenasen, in diese Gruppe fallen auch die Lactat-Dehydrogenase (LDH), die Leber-Alkohol-Dehydrogenase (LADH) und die Glyceraldehyde-3-phosphate Dehydrogenase. Alle diese Enyzme zeigen eine hohe Faltungsähnlichkeit, trotzdem wird eine sehr hohe Substratselektivität beobachtet [Goward and Nicholls, 1994]. Beispielsweise setzt cMDH Malat und Oxalacetat um sechs Größenordnungen effizienter als Pyruvat und Laktat um. Eine bewegliche Schleife nahe der Substrat-Bindestelle ist für die unterschiedliche Substratspezifität verantwortlich. Ein struktureller Vergleich eines cMDH (1mld) Binärkomplex mit einem cMDH Tertiärkomplex (5mdh) verdeutlicht die große Beweglichkeit der Schleifen nahe des aktiven Zentrums, besonders das katalytische Arg97 nimmt unterschiedliche Konformationen ein. Die von Cavbase detektierten Bindetaschen sind an der Schnittstelle zwischen den verschiedenen Untereinheiten lokalisiert. Sie formen im Zentrum eine sehr große Bindetasche. Wie durch Ligsite [Hendlich et al., 1997] ausgeschnitten umfasst sie alle Reste der vier katalytischen Zentren aus jeder Untereinheit (Abbildung 4.21). Die Kristalle für die Strukturbestimmung (PDB Code 1mld) wurden aus einem Citrat-haltigen Puffer gewonnen, daher ist in der Bindetasche ein Citratmolekül gebunden. Die Bindestelle wird durch die Aminosäuren Arg80, Arg86, Arg152, His176 and Asp149 (Aminosäurenummerierung wie in 1mld) geformt und ist in der Lage, das negative geladene Citratmolekül aufzunehmen (siehe Abbildung 4.22). Um mögliche Ähnlichkeiten in der Bindetasche der MDH und CA Enzymfamilien zu entdecken, wurde ein Datensatz aus 24 Carbonanhydrasen aller strukturell bekannten Unterfamilien gegen einen Satz von 71 Malatdehydrogenasen verglichen (siehe Tabelle 4.11). Neben den bereits oben erwähnten MDH-Typen (cMDH und mMDH) sind noch weitere Malatdehydrogenasen, wie Malatdehydrogenasen aus *E. coli* (eMDH) oder Malatdehydrogenasen aus thermophilen Organismen (tMDH, caMDH, ctMDH, etc.), im Datensatz vertreten. In diesem



(**I**) (**II**) (**III**)

Abb. 4.21 Kristallstruktur und Faltungsmuster der Malatdehydrogenase (1mld). In (I) ist das Faltungsmuster und die sehr große Bindetasche von 1mld gezeigt, die alle vier Monomere umfaßt. (II) zeigt die Aminosäuren, die Pseudozentren und die gebundenen Citratmoleküle (gelb gefärbt) und in (III) ist die Bindetaschenoberfläche dargestellt.

Tab. 4.9 Übersicht der katalytische Bindetaschen der Malatdehydrogenase.

Cavbase Kennung
5mdh.1 1bmd.1 1bmd.7 1mld.1 1emd.1 1guz.1 1guz.2 1guz.3 1gv0.2
1gv0.5 1gv1.1 1ib6.2 1ib6.5 1ie3.1 1ie3.3 2cmd.1 4mdh.1

Datensatz sind neben katalytischen MDH Bindetaschen (siehe Tabelle 4.9) auch nicht katalytische Bindetaschen (siehe Tabelle 4.10) enthalten. Die Einteilung in katalytische und nicht-katalytische Bindetasche wurde nach visueller Beurteilung der Malatdehydrogenasen vorgenommen und hat sich am Vorhandensein der katalytischen Reste in der Bindetasche orientiert.

Tab. 4.10 Übersicht über weitere Bindetaschen in MDHs, die nicht die katalytischen Reste umfassen.

Cavbase Kennung
1bmd.2 1bmd.3 1bmd.4 1bmd.5 1bmd.6 1bmd.8 1d3a.1 1d3a.2 1d3a.3
1d3a.4 1d3a.5 1d3a.6 1emd.2 1guy.1 1guy.2 1guy.3 1guy.4 1guy.5
1gv0.1 1gv0.3 1gv0.4 1gv1.2 1gv1.3 1ib6.1 1ib6.3 1ib6.4 1ib6.6
1ib6.7 1ie3.2 1ie3.4 1ie3.5 1ie3.6 1ie3.7 1ie3.8 1mld.2 1mld.3
1mld.4 1mld.5 1mld.6 1mld.10 1mld.7 1mld.8 1mld.9 2cmd.2 2hlp.1
2hlp.2 2hlp.3 2hlp.4 2hlp.5 4mdh.2 4mdh.3 4mdh.4 5mdh.2 5mdh.3

4.4.3 Ähnlichkeitssuche und Analyse der Ergebnisse

Aufgrund der enormen Ausdehnung und des großen Volumens der berücksichtigten Malatdehydrogenase-Bindetaschen wurden die Einstellungen für den Vergleichsalgorithmus so gewählt, daß sie einen größeren Suchraum abdecken. Für jeden eins-zu-eins Vergleich wurden 10000 Lösungen generiert, die beste Überlagerung aus dem Clique-Algorithmus wurde anhand der Größe der überlagerten Bindetaschenoberflächen bestimmt. Die Vergleiche wurden mit dem Original Cavbase Ansatz durchgeführt.

Cavbase erzielt eine klare Separation von katalytischen und nicht-katalytische Bindetaschen der MDHs. Obwohl die MDH-Bindetasche sehr groß ist, findet Cavbase erstaunlicherweise die größte Ähnlichkeit genau im Bereich der katalytisch wichtigen Reste in beiden aktiven Zentren. Die in beiden Fällen für die Katalyse wichtigen Aminosäuren werden zur Übereinstimmung gebracht. Auch die gebundenen Liganden/Inhibitoren überlagern räumlich (siehe Abbildung 4.24). Für die katalytische Reaktion der MDH sind die Aminosäuren Arg91, Arg97, Asp158, Arg161, His186 und Ser241 (Abbildung 4.22) sehr wichtig, in der Carbonanhydrase sind die Reste His94, His96, His119 und Thr199 in die Zinkkoordination und Inhibitorbindung involviert. In Tabelle 4.12 sind die vom Vergleichsalgorithmus als ähnlich erkannten Pseudozentren und Aminosäuren gezeigt. Ähnliche physikochemische Eigenschaften in den Bindetaschen der Malatdehydrogenase und den Carboanhydrasen sind möglicherweise für die beobachtete Kreuzreaktivität von E7070 verantwortlich. Gewissheit kann hier allerdings nur die Aufklärung der räumlichen Struktur einer cMDH in Komplex mit gebundenem E7070 geben. Zur Zeit wird in einer Kooperation mit der Arbeitsgruppe von Prof. C. Supuran (Universität Florenz) versucht, die Kristallisation von E7070 mit cMDH durchzuführen. Die

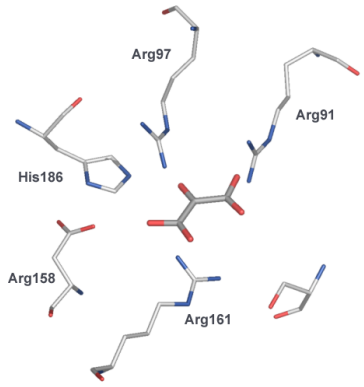


Abb. 4.22 Katalytische Reste in der Bindetasche der cMDH von (5mdh).

als ähnlich entdeckten Bereiche zwischen den katalytisch entscheidenden Aminosäuren beider Enzyme dienen als ein erster Hinweis, daß E7070 gegebenenfalls in das aktive Zentrum der cMDH binden könnte. Die Bindungsexperimente von Oda und Kollegen zeigen, daß E7070 sowohl in der freien Form mit terminaler Sulfonamidgruppe, als auch mit nur einer Sulfonamidgruppe und Fixierung an den Linker an die cMDH bindet. In welcher Konformation E7070 allerdings an cMDH bindet, ist nicht genau abzuschätzen.

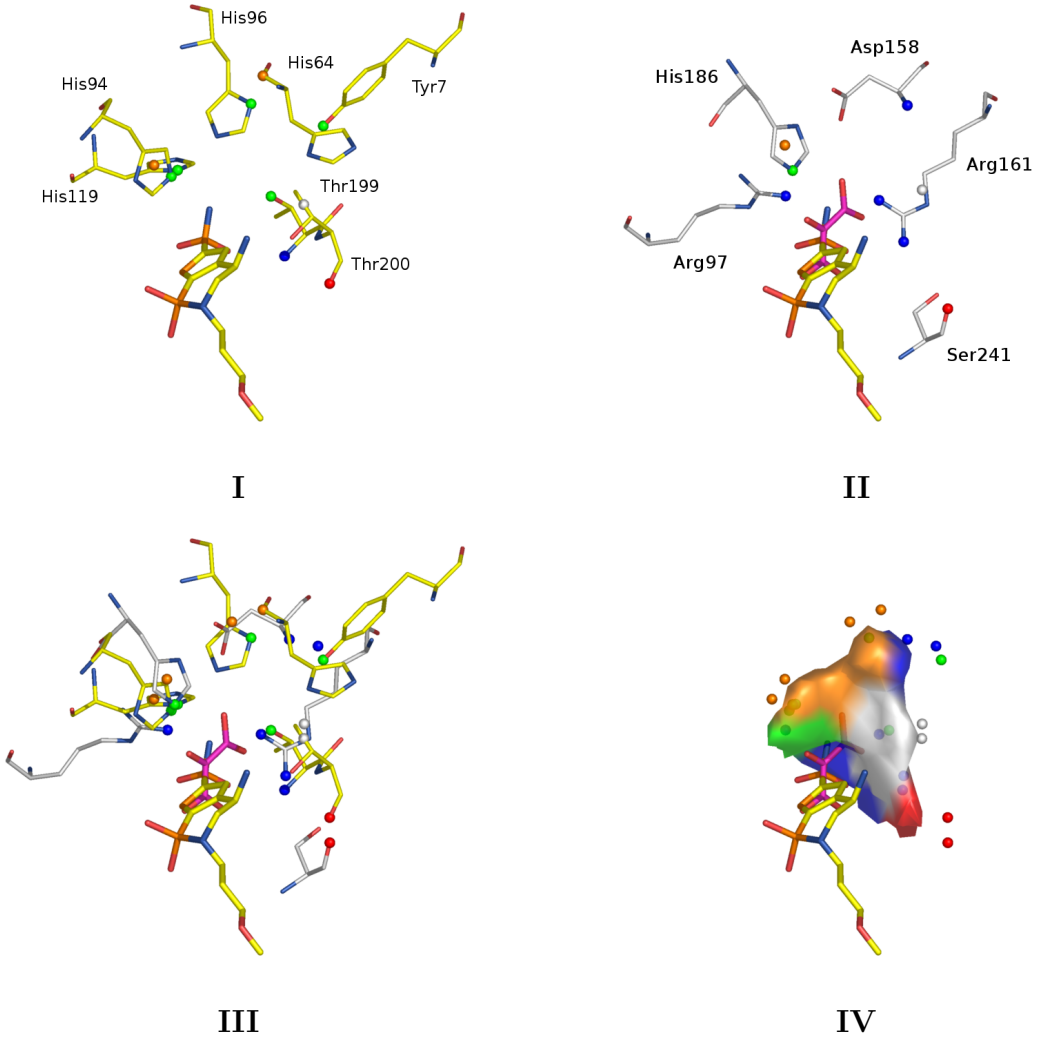


Abb. 4.23 Bereiche aus den Bindetaschen der CA II (PDB Code 1i90) und cMDH (PDB Code 5mdh), die vom Vergleichsverfahren als gemeinsames Muster erkannt werden. In (I) sind die Aminosäuren und Pseudozentren der Carboanyhdrase, die als ähnlich zur MDH angesehen werden, dargestellt (Kohlenstoffe in gelb gefärbt). In (II) sind die Aminosäuren und die Pseudozentren der MDH mit gebundenem α-Keto-Malonat, die als ähnlich zur CA gefunden werden, gezeigt (Kohlenstoffe in weiß). Zusätzlich ist der CA-II Inhibitor Al-8520 (**11**, Kohlenstoffe in gelb) in der gefundenen Überlagerung dargestellt. (III) und (IV) zeigen die Überlagerung der beiden Taschen, (IV) die gemeinsamen Pseudozentren und die ähnlichen Oberflächenbereiche (hier nur gezeigt für die CA-II 1i90 Tasche). Das α-Keto-Malonat (violett, (**12**)) fällt in den gleichen Bereich, der auch vom CA-Inhibitor Al-8520 besetzt wird.

Tab. 4.11 In der Vergleichsanalyse verwendete Proteinstrukturen der Malatdehydrogenasen.

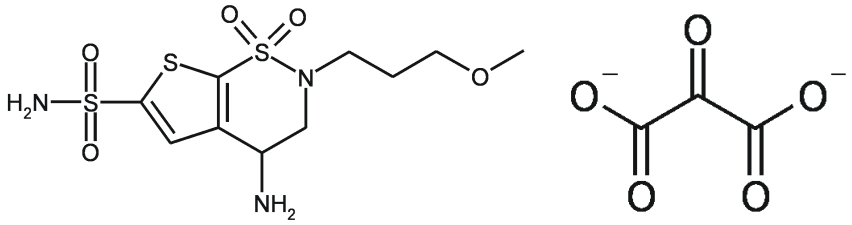
PDB-ID	MDH-Typ	kat. Ratio[1]	Anmerkungen
5mdh	cMDH	1/3	Dimer, Tertiärkomplex mit α-Ketomalonat, Tetrahydro-NAD große Bindetasche
1mld	mMDH	1/10	Tetramer, gebundenes Citrat, sehr große Bindetasche
1bmd	tMDH	2/8	Dimer, Binärkomplex mit NADH
1d3a	HmMDH	0/6	MDH aus hitzestabilem Organismus, nur Ionen gebundenen, keine typische MDH-Bindetasche
1emd	eMDH	1/2	Ternärkomplex mit gebundenem Citrat und NAD
1guy	caMDH	2/5	Binärkomplex mit NAD, die katalytische Reste sind auf mehrere Taschen verteilt
1guz	ctMDH	3/3	Binärkomplex mit NAD, sehr große Bindetasche
1gv1	cvMDH	1/3	kein Ligand gebunden, sehr große Bindetasche
1ib6	eMDH	2/7	Binärkomplex mit NAD und Sulfation, Mutantenstruktur
1ie3	eMDH	2/8	Ternärkomplex mit gebundenem Pyruvat und NAD (eine Tasche hat nur NAD gebunden)
2cmd	eMDH	1/2	kleine Bindetasche
2hlp	HmMDH	0/6	MDH aus hitzestabilem Organismus, nur Ionen gebundenen, keine typische MDH-Bindetasche
4mdh	cMDH	1/3	Binärkomplex mit gebundenem Citrat

[1] Verhältnis von katalytischen zu nicht katalytischen Bindetaschen für das MDH Protein.

Tab. 4.12 Ähnliche Reste in den Bindetaschen von MDH (PDB Code 4mdh und 5mdh) und CA II (PDB Code 1i90). In beiden Strukturen werden die katalytisch wichtigen Reste zur Übereinstimmung gebracht

Carboanhydrase (1i90.1)			**cMDH** (4mdh.1)		
Typ des Pseudo-zentrums	äquivalente Aminosäure[a]		Typ[a]	äquivalente Aminosäure[a]	
Donor-Akzeptor	T200	s	Donor-Akzeptor	H186	s
Donor-Akzeptor	H119	s	Donor-Akzeptor	S241	s
Aromatisch	H94	s	Pi	G230	p
Donor	N67	s	Donor	A231	p
Aromatisch	H64	s	Pi	S187	p
Donor-Akzeptor	H64	s	Donor	S188	p
Akzeptor	H64	p	Akzeptor	V226	p
Pi	H64	p	Pi	Q227	p
Akzeptor	N62	p	Donor-Akzeptor	S188	s
Donor-Akzeptor	Y7	s	Akzeptor	D158	s
Carboanhydrase (1i90.1)			**cMDH** (5mdh.1)		
Typ des Pseudo-zentrums	äquivalente Aminosäure[a]		Typ des Pseudo-zentrums	äquivalente Aminosäure[a]	
Aliphatisch	T 200	s	Aliphatisch	L 157	s
Akzeptor	T 200	p	Akzeptor	S 241	p
Donor-Akzeptor	T 199	s	Donor	R 161	s
Donor	T 199	p	Donor	R 161	s
Donor-Akzeptor	H 119	s	Donor	R 97	s
Donor-Akzeptor	H 96	s	Donor	D 158	p
Aromatisch	H 94	s	Aromatisch	H 186	s
Donor-Akzeptor	H 94	s	Donor-Akzeptor	H 186	s
Pi	H 64	p	Pi	L 154	p
Donor-Akzeptor	Y 7	s	Donor	L 157	p

[a] Ein-Buchstabencode der Aminosäure, Positionsnummer und Lokalisation des Pseudozentrums: auf der Seitenkette (s) oder auf der Peptidbindung (p).



11 12

Abb. 4.24 Chemische Formeln von dem CA-II-Inhibitor Al-8520 (11) und α-Ketomalonat (12).

4.5 Zusammenfassung und Schlussfolgerungen

Eine große Herausforderung an Verfahren, die Proteine anhand struktureller Ähnlichkeiten und Eigenschaftsmerkmalen der Bindestelle vergleichen, stellt das Entdecken unerwarteter Ähnlichkeiten zwischen nicht-verwandten Proteinen dar. Zunächst aber muß ein solches Verfahren in der Lage sein, bekannte Ähnlichkeiten verwandter Proteine bei Datenbanksuchen in großen Datensätzen sicher zu identifizieren. In Abschnitt 4.1.1 und Abschnitt 4.1.3 wurde die erfolgreiche Suche nach Proteinen, die einem ähnlichen katalytischen Mechanismus folgen, vorgestellt. Cavbase ist in der Lage, funktionelle Ähnlichkeiten in der Bindetasche sicher zu erkennen und diese auch nach abgestufter Ähnlichkeit der Proteine relativ zueinander zu bewerten. Proteine, die einen ähnlichen katalytischen Mechanismus aufweisen, werden klar von solchen separiert, die eine andere Funktion ausüben.

Der Einsatz von Cavbase in der Analyse einer strukturell diversen Gruppe von Proteinen, die alle einen bestimmten Kofaktor binden, wurde in Abschnitt 4.1.2 vorgestellt. Die untersuchten Proteine im Datensatz unterscheiden sich in Sequenz und Faltung. Cavbase ist hier in der Lage, das wiederkehrende Motiv der NAD(P)-Bindung zu erkennen, wichtige Interaktionen des Proteins mit dem Kofaktor werden wiederholt gefunden. Auch ist es mit Cavbase erfolgreich möglich, Bindestellen verwandter Kofaktoren auf den folgenden Rängen zu detektieren. Interessanterweise werden auch solche Beispiele als ähnlich entdeckt, bei denen sich die Kofaktorbindung aus sequenziell komplett unterschiedlichen Aminosäuren zusammensetzt. Gerade dies zeigt die Stärke und das Potential von Ansätzen, die, ohne auf Sequenz- und Faltungsinformationen angewiesen zu sein, Ähnlichkeiten in Bindetaschen detektieren.

In Kapitel 4.2 wurden besonders interessante und herausfordernde Beispiele für die Funktionsannotation von Proteinen vorgestellt: Proteine bekannter Struktur aber noch unbekannter Funktion dienten als Testfälle für eine mögliche Funktionsannotierung durch Vergleiche mit Einträgen in der Datenbank. Cavbase ist in der Lage, funktionelle Ähnlichkeiten mit anderen Proteinen zu entdecken und Ideen für die Funktionsannotierung vorzuschlagen. Sind zu dem Protein unbekannter Funktion verwandte Proteine strukturell charakterisiert, dann ist es mit Cavbase möglich, diese ähnlichen Bereiche in den Bindetaschen zu detektieren. Eine besondere Herausforderung an die Funktionszuweisung ist die Annotation von Strukturen, für die nur wenig verwandte Strukturen bekannt sind. In diesen Fällen ist es möglich, mit Cavbase Informationen über Liganden

oder Ligandfragmente zu sammeln, die in ähnliche Subtaschen binden. Eine Voraussetzung für dieses Vorgehen ist aber, daß die zum Vergleich herangezogenen Bereiche in Vertiefungen auf der Proteinoberfläche liegen und somit als Teil einer Bindetasche detektiert werden.

Das Kapitel schließt mit zwei Beispielen für die Suche nach Kreuzreaktivitäten zwischen Proteinen. Im ersten Fall (siehe Kapitel 4.3) konnte eine im Experiment beobachtete Kreuzreaktivität strukturell verstanden werden und Cavbase detektiert ähnliche Bereiche in den Bindetaschen. Im zweiten Beispiel wurden ähnliche Bereiche in den Bindetaschen von Carboanhydrase und Malatdehydrogenase aufgefunden, die eine im Experiment beobachtbare Kreuzreaktivität strukturell plausibel machen (siehe Kapitel 4.4.1). Den experimentellen Beweis hierfür muß eine Kristallstrukturbestimmung erbringen. Cavbase hat das Potential, Kreuzreaktivitäten zwischen nicht verwandten Proteinfamilien zu entdecken. Die Wahrscheinlichkeit solche Übereinstimmungen zu finden, wird besonders dadurch erhöht, daß man Subtaschen für die Ähnlichkeitsanalyse benutzt und den Grad der dort entdeckten Ähnlichkeiten kombiniert. So ist man in der Lage, Ähnlichkeiten zu finden, die man bei Vergleichen kompletter Bindetaschen nicht entdecken würde.

5 Functional classification of protein families

In this chapter, the classification of protein families based on the properties of their active sites is presented. Most methods that analyze similarities in protein structures compare one query structure of interest against a structural database (see chapter 2). Here, the clustering analysis of complete protein families will be presented. The chapter is organized as follows: section 5.1 will shortly introduce the Cavbase method, followed by a detailed description of the clustering procedure 5.2. In section 5.3 and 5.4 two application scenarios using pharmaceutically relevant protein families will be presented: the carbonic anhydrases and the eukaryotic protein kinases.

5.1 Cavbase - a method to describe and compare protein binding pockets

Cavbase is a method to describe and compare protein binding pockets in terms of exposed physicochemical properties [Schmitt et al., 2001, 2002]. Binding pockets are automatically extracted from proteins using the Ligsite algorithm [Hendlich et al., 1997] and stored in Cavbase. Cavbase is integrated and hyperlinked to the protein-ligand database Relibase [Hendlich, 1998; Gunther et al., 2003; Hendlich et al., 2003]. In the current version Cavbase used for this analysis, information about 80661 binding pockets extracted from 22885 proteins is stored.

In Ligsite, the protein under consideration is embedded into a regularly-spaced Cartesian grid with 0.5Å grid spacing. Any grid points, represented by 1.5Å probe spheres, penetrating into protein atoms based on their van der Waals radius are discarded as solvent-inaccessible grid points. The remaining ones are classified in terms of their degree of burial in the protein binding pocket. Grid points with a high degree of burial are merged together forming contiguous cavities. All surface-contacting grid points of such a cluster, apart from the non-buried ones oriented towards the solvent, are used to approximate the cavity surface. If one atom of an amino acid residue falls closer than 1.1 Å to a protein surface-contacting grid point, the amino acid is classified as a

cavity-flanking residue. These data are used to represent the basic geometric shape of cavities in the database [Schmitt et al., 2002].

The physicochemical properties of the amino acids flanking the cavity are encoded in terms of pseudocenters that represent appropriate 3D descriptors. These pseudocenters are coordinates in 3D space associated with properties determinant for molecular recognition: hydrogen-bond (HB) donor, HB acceptor, mixed HB donor/acceptor, hydrophobic aliphatic contact, and aromatic contact (Figure 5.1). This condensed representation allows for efficient similarity searches on the basis of a reduced set of input variables. The property of the assigned pseudocenters is projected onto a particular area of the cavity surface, in order to verify whether a particular interaction property can form an interaction to a potential ligand. Only those pseudocenters, that are able to project their property onto the cavity surface, are retained. The amino acids flanking the binding site, the grid points with their degree of burial, the pseudocenters, as well as the attributed cavity surface points are stored in Cavbase [Schmitt et al., 2002].

For the comparison of different binding sites, a clique algorithm [Bron and Kerbosch, 1973] is applied to detect common substructures within two cavities. Two binding sites are regarded as similar if they share a common spatial arrangement of assigned pseudocenters and expose similar physicochemical properties into the binding site. The clique detection handles the pseudocenters as nodes of a graph considering the assigned physicochemical property. An edge is introduced between two nodes and weighted according to their mutual geometric distance. A standard graph algorithm [Bron and Kerbosch, 1973] is used to detect connected common substructures in the graph representations. For each mutual comparison of cavities, a predefined number of clique solutions is considered. Each detected clique solution is subsequently scored according to the degree of spatial overlap in equivalent cavity surface patches. These surface patches originate from the embedded grid of 0.5 Å spacing. Three different scoring schemes are used to rank the obtained superpositions. The degree of mutual overlap in surface patches is expressed by the number of surface points attributed to one of the five properties that fall next to each other below a distance threshold of 1.0 Å. To avoid consideration of weakly overlapping surface patches, the mutual overlap of two patches is only counted if at least 70% of the matched surface patches, corresponding to a matched pseudocenter pair, fall next to each other below 1.0 Å. Scoring scheme R_1 is thus calculated by adding the percentage of overlapping surface points from both pseudocenters divided by the total number of surface points of these patches in agreement with these criteria. This score depends linearly on the size of the matching cavity surfaces consi-

dered in both structures. Scoring scheme R_2 (Equation (5.1)) reflects in addition the root-mean-square deviations (rmsd) of the detected matching pseudocenters (n) and downranks fragmented non-contiguous clique solutions that obtain an artificially high R_1 score. Scoring scheme R_3 (Equation (5.1)) is calculated analogous to the Tanimoto index [Godden et al., 2000]. It accounts for the number of commonly matched pseudocenters n_{ctr} in both cavities but normalizes this value with respect to the entire size of the cavities expressed by the total number of pseudocenters $n_{maxcav1}$ and $n_{maxcav2}$, respectively.

$$\begin{aligned} R_2 &= \frac{R_1 - 0.7\ n}{rmsd} \\ R_3 &= \frac{n_{ctr}}{n_{maxcav1} + n_{maxcav2} - n_{ctr}} \end{aligned} \tag{5.1}$$

This score is further normalized to values between zero and one and adopts a value of one, if a cavity is matched upon itself.

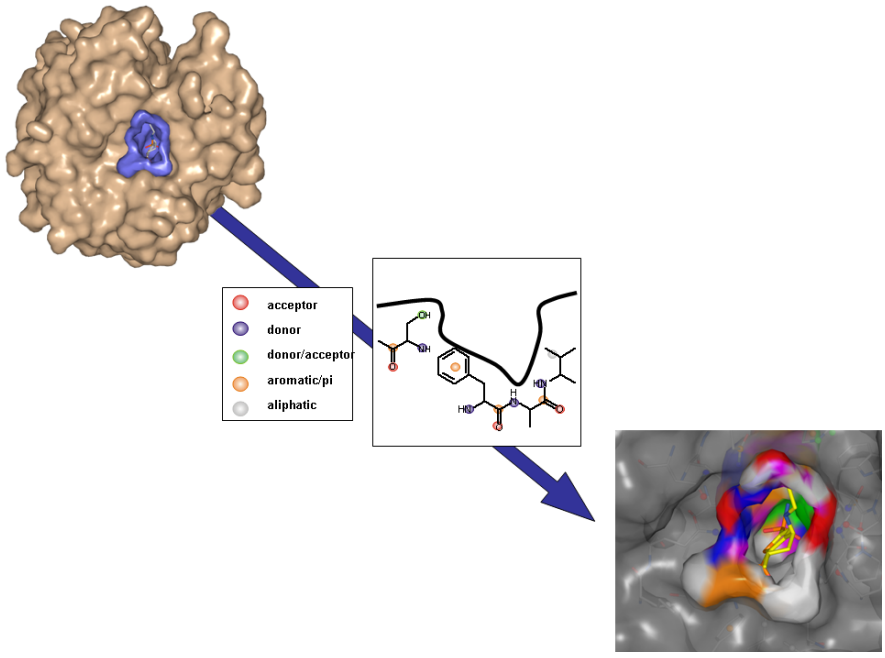


Fig. 5.1 From protein structure to binding site. Cavities are detected as depressions on the protein surface. Physicochemical interaction properties of the amino acids flanking the cavity are encoded into 3D descriptors in terms of assigned pseudocenters. Pseudocenters are displayed as colored spheres. Color scheme: H-bond donor (blue), H-bond acceptor (red), ambivalent donor/acceptor (green), hydrophobic aliphatic (white), aromatic (orange). Only those pseudocenters are retained that expose their property onto the cavity surface. Binding pockets classified by this way are stored in Cavbase.

5.2 Cavity clustering procedure

In an initial clustering analysis 113 cavities from 14 functional diverse protein families were used to validate and examine the anticipated clustering procedure. All members of the protein families were enzymes. Their EC-annotation was used to assess the quality of the obtained clustering solutions. Existing classification schemes for proteins such as SCOP [Lo Conte et al., 2002], CATH (structural and functional relationships) [Pearl et al., 2000] or the ENZYME [Bairoch, 2000] database (functional annotation of protein structures) provide valuable information to benchmark the classification results obtained with Cavbase. The Cavbase clustering analysis was able to clearly distinguish between the 14 different protein families. Furthermore, existing relationships between the different protein families could be detected. For example, similarities between related protein subfamilies such as serine proteases of the subtilisin-like and trypsin-like protein family or protein kinases from the tyrosine and serine/threonine class, could be found (results not shown). After these encouraging results the applicability of the approach to the classification of homologous protein families was investigated: the α-carbonic anhydrases and the eukaryotic protein kinases (see section 5.3 and 5.4).

Each cavity in the dataset is mutually compared to all other cavities. For each one-to-one comparison Cavbase returns three similarity scores depending on the applied scoring scheme. The resulting scores are stored in a similarity matrix, measuring the similarity across the considered cavities. This similarity matrix serves as input for various clustering algorithms. The clustering is performed here using the clustering toolkit Cluto [Karypis, 2002], which provides a variety of clustering algorithms. Optionally the clustering of the cavities can be performed according to an agglomerative, a graph-partitioning, and two partitional clustering strategies [Jain et al., 1999]. Since the number of anticipated cluster solutions has to be defined prior to clustering, a variety of parameter settings were tested and empirically validated. As a general rule it is advisable to provide a number in the range of the expected subfamilies, which are present in the dataset[1]. An optimal similarity measure would have the following properties: reflexive (cavity A is similar to itself), symmetric (cavity A is similar to cavity B and vice versa), transitive (cavity A is similar to B and cavity B is similar to cavity C; cavity A is therefore similar to cavity C). Whereas for the Cavbase score the first two conditions are met, especially the transitive term is not necessarily fulfilled.

[1]For a detailed analysis of the carbonic anhydrases and the kinases a number of 6 and 48 clusters respectively has proven to be useful.

Although the Cavbase similarity scores do not fulfill the requirements of a metric, the obtained clustering solutions are rather insensitive to different parameter settings and the individual scoring schemes applied, accordingly they all produce consistent results. They were visually inspected using a 2D-plot (Figure 5.10). The intensity of the reddish coloring indicates the degree of mutual similarity. Along the principal axis the individual clusters are plotted, separated by solid black lines (Figure 5.10). The grouping into clusters is performed by the algorithm solely based on the similarity matrix as input. In total twelve different combinations of clustering options were evaluated, combining four different clustering strategies with three different scoring schemes.

5.3 α-carbonic anhydrases

The carbonic anhydrase (CA) gene family comprises at the present state pf research fourteen active members. Their basic physiological functions is linked to the interconversion of carbon dioxide and bicarbonate ($CO_2 + H_2O \rightleftharpoons H^+ + HCO_3^- \rightleftharpoons H_2CO_3$). They participate in a variety of physiological processes that involve pH regulation, CO_2 and HCO_3^- transport as well as water and electrolyte balance [Lindskog, 1997]. Three major families of carbonic anhydrases exist: α-, β-, and γ- carbonic anhydrases, which are unrelated by sequence. All members of these three different families catalyze the same reaction and share a similar zinc binding region in common [Lindskog, 1997]. Carbonic anhydrases from the animal kingdom are all of the α-type. These different CA-α isozymes possess a similar protein architecture consisting of a 10-stranded twisted beta sheet. However, they show different levels of sequence identity. The active site is formed by a large cone-shaped cavity with a zinc ion at the bottom. The zinc ion is tetrahedrally coordinated by three histidines and most likely due to a pk_a shift a hydroxide ion. The residues involved in the zinc binding are invariant, additionally 17 residues are also found across all carbonic anhydrase structures [Lindskog, 1997]. Crystal structures are available for six of the 14 isozymes of the carbonic anhydrases (Fig. 5.2). According to the SCOP database, there are 173 carbonic anhydrase structu-

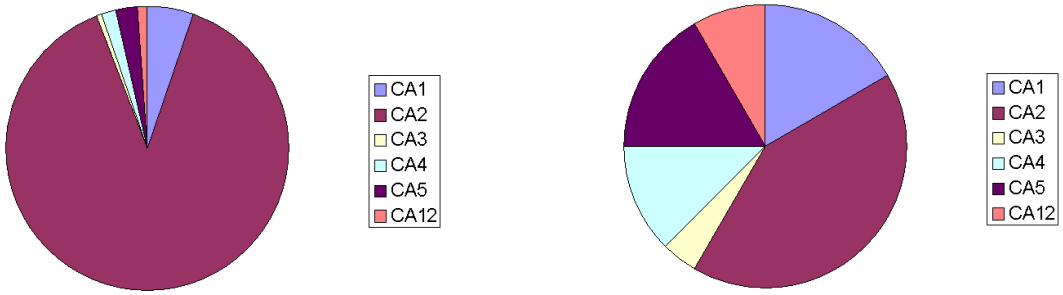


Fig. 5.2 (I) **Distribution of crystallopraphically determined carbonic anhydrase structures [PDB version June 2003] (II) Distribution of the carbonic anhydrases used as dataset**

res deposited in the PDB belonging to the superfamily of α carbonic anhydrases. The majority of these structures belongs to the class of the carbonic anhydrases II. Table 5.1 summarizes the 24 cavities considered in the present analysis and Fig. 5.2 shows the distribution of the different carbonic anhydrase family enzymes across the considered dataset. All structures from the less populated carbonic anhydrase families were selected for the present cluster analysis. Only cavities that accommodate the catalytic

residues coordinating the zinc ion were retained. A consistent classification is obtained consulting the SCOP and ENZYME databases together with general knowledge about the different CAs isozymes.

Tab. 5.1 Manually determined clustering of CAs. The clustering was obtained using structural information and knowledge from the literature. This scheme serves as a guide for the evaluation of different clustering parameters. The CA-IIs are distinguished in two classes, depending on the conformation of a catalytic important residue.

Carbonic anhydrase	Cavbase ID
CA-I	1azm.1 1bzm.1 1czm.2 1hcb.1
CA-IIa	1cil.1 1g52.1 1g54.1 1i8z.1 1i90.1 1a42.1 1if4.1 1if8.1
CA-IIb	1bcd.1 1ca2.1
CA-III	1flj.1
CA-IV	1znc.3 2znc.2 3znc.2
CA-V	1dmx.1 1dmy.1 1urt.1 1keq.1
CA-XII	1jcz.1 1jd0.1

5.3.1 Carbonic anhydrase classification results

The classification of the CAs shows that best clustering results were obtained using a combination of scoring scheme R_1 and clustering algorithm rb. Nevertheless, the clustering results based on the other scoring functions (R_2 , R_3) and different algorithms during clustering (rbr,agglo) reveal consistent classifications. For example the classification based on four different combinations are displayed in Fig. 5.3. Cavbase is able to cluster the CAs convincingly on a subfamily level and distinguishes even between members of different CA isozymes. Interestingly, cavities from the CA-II are grouped in two distinct clusters. Further structural analysis revealed that this separation originates from different conformers among the CA-II entries. Cavities from both clusters differ in particular with respect to the conformation of His64. It is known that this

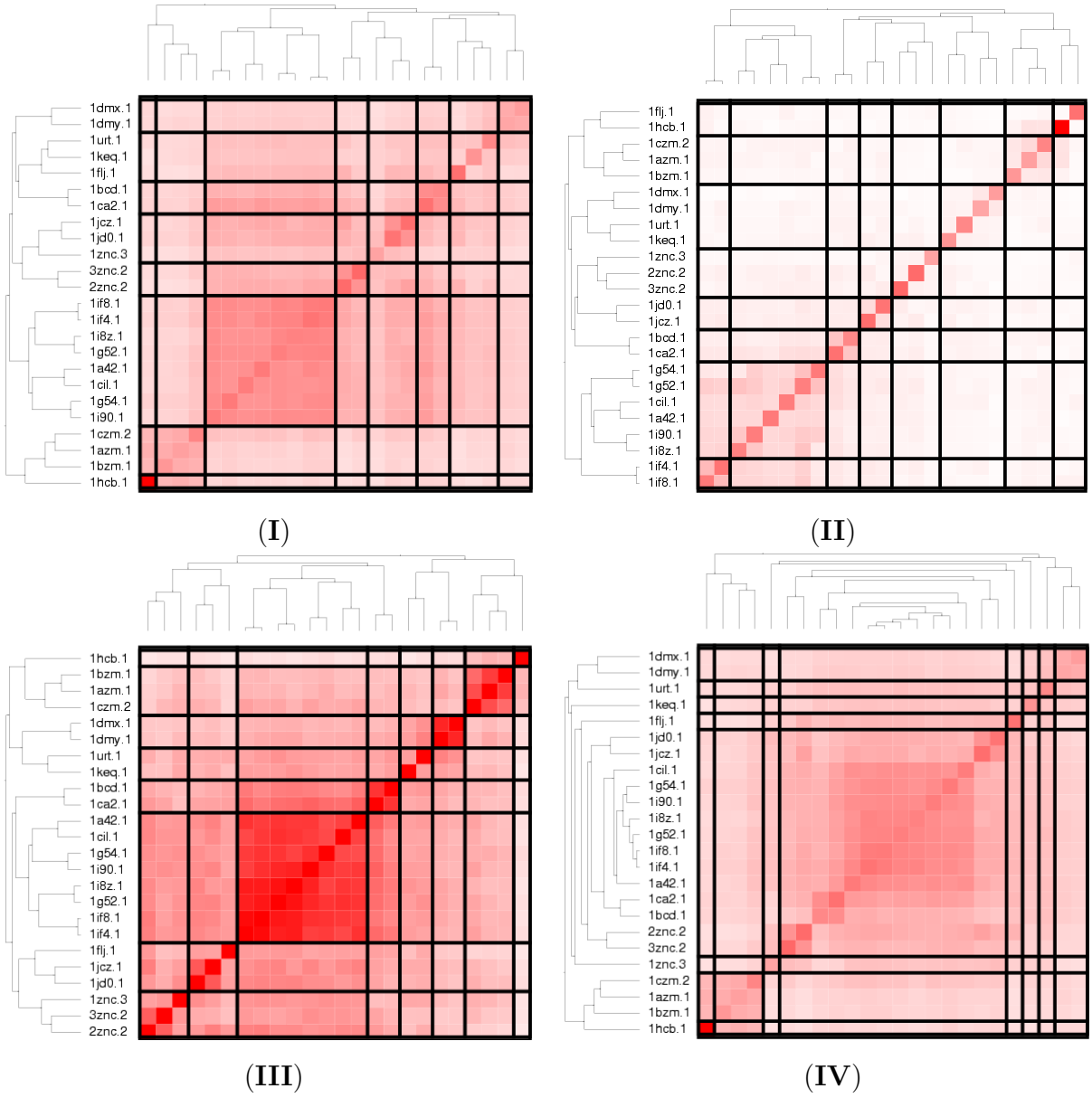


Fig. 5.3 Clustering results for the α-CA isozymes. The influence of different clustering parameters is presented for the example of CA classification. In all presented examples eight clusters were obtained. In (I) to (III) the clustering algorithm rb was used in conjunction with the scoring schemes R_1, R_2, and R_3 respectively. The different scoring schemes give consistent results and provide reasonable clusterings. The clustering obtained using R_2 separates CA-II cavities exhibiting the His64 in-conformation into two clusters. Differences between these cavities can be observed at the rim of the binding pocket, where the pocket opens towards the solvent. In (IV) the result for the usage of clustering algorithm agglo in combination with scoring scheme R_1 is shown. The clustering function agglo tends to separate singletons early and produces often one large cluster. As a matter of fact, using a higher number of clusters would separate the different entries of that cluster.

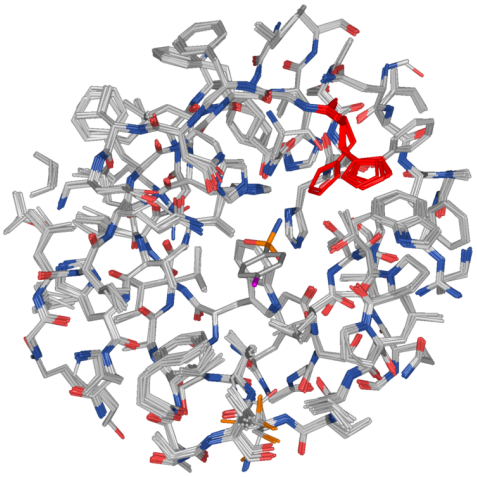


Fig. 5.4 Superposition of CA-II structures showing the flexibility of His64. The superposition was obtained using Relibase. Cavbase distinguishes between CA-II structures that differ in the conformation of His64.

residue is flexible and plays an important role in the catalytic mechanism of carbonic anhydrases, acting as a proton shuttle [Kim et al., 2002]. His64 can adopt a distinct 'in' and 'out' conformation, and Cavbase distinguishes between the CA-II structures with different conformations of His64. In Table 5.1 both CA-II isozymes corresponding to two different conformations are labelled either CA-IIa ('in' conformation) or CA-IIb ('out' conformation). Fig. 5.4 shows the active site regions aligned with Relibase+.

Four CA-I cavities are considered in the dataset, one of them (PDB code 1hcb) performs differently from the other three. Obviously, it shows only low similarity to other carbonic anhydrases in the clustering analysis. The automatically extracted cavity for this protein entry is considerably larger than the other three, accordingly it yields a very high self-similarity score. Nevertheless, all four combinations of clustering parameters are able to capture this cavity as a member of the CA-I cluster (Fig. 5.3).

The structure of CA-IV shows a general similarity to other CA isozymes (e.g. CA-II), however there are some differences. Most notably, the residues from Val131 to Asp136 adopt in CA-IV a loop conformation pointing towards the solvent, whereas in other CA isozymes an α-helix conformation is found. This helix is directed towards the binding site. A comparison of a CA-II and CA-IV cavity with Cavbase reveals no similarity in that region. Three CA-IV cavities are used in the present dataset. The human CA-IV structure (PDB code 1znc) shows a sequence identity of 56% to the two murine CA-IV structures (PDB code 2znc and 3znc). Several amino acid substitutions are found in the active site (e.g. K91V, M67E, S65T, I141F) among the cavities from the different

species. Furthermore, the cavity 1znc is significantly smaller than the other two cavities. Despite these differences, scoring scheme R_2 and R_3 are able to classify all three CA-IV cavities into one cluster (Fig. 5.3).

Cavbase detects the similarity between all four CA-V cavities. However, depending on the clustering parameters used, the CA-V cavities are found in a different number of subclusters (Fig. 5.3). There are two CA-V wild-type (1dmx and 1dmy) and two double mutant (1keq (F65A/Y131C) and 1urt (Y64H/Y131A)) isozymes in the present dataset. The similar areas between both wild-type (Fig. 5.5-I) and the wild-type and a mutant cavity are shown (Fig. 5.5-II). The different physicochemical properties of the mutated amino acids cannot be matched. Nevertheless, all CA-V exhibit such a degree of similarity in the binding sites to be clustered together.

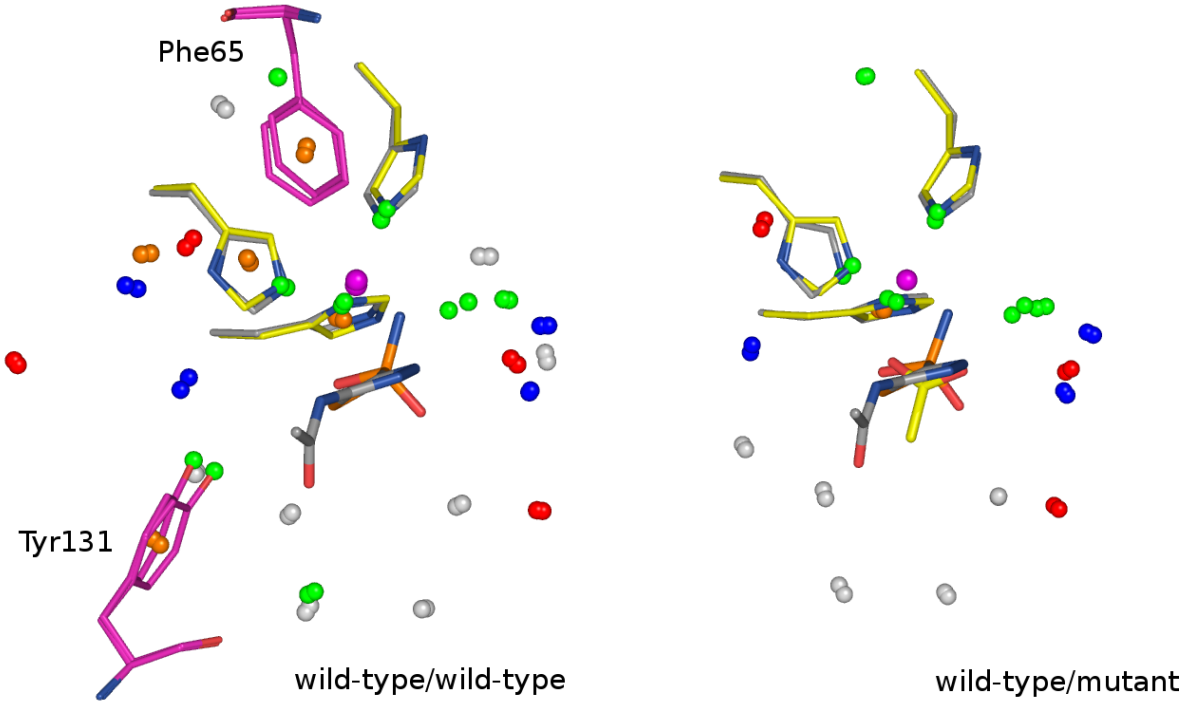


Fig. 5.5 Similar areas in the binding sites of CA-V wild-type (PDB code 1dmx and 1dmy) and a mutant isozyme (PDB code 1keq). In (I) the similar areas in the binding sites of both CA-V wild-type cavities are shown and (II) displays the similar areas in the binding sites of a wild-type cavity (PDB code 1dmy) compared to a CA-V mutant. For reasons of clarity only the matching pseudocenters, the bound zinc ions and the sulfonamide inhibitor (1dmy) together with the three histidines involved in zinc binding are shown. Additionally, in (I) the phenylalanine and tyrosine are displayed (carbon atoms in magenta), which are mutated to alanine and cysteine, respectively. These changes in the physicochemical properties of the amino acids cannot be matched.

5.3.2 Sequence-based classification of carbonic anhydrases

Furthermore for reasons of comparison, a clustering was performed using sequence identity values as similarity measure. Sequence identity was calculated using the FASTA program, version 3.4 [Pearson and Lipman, 1988] for only those protein chains, that are involved in formation of the cavity. Fig. 5.6 shows the different clustering results. Sequence similarity measurements are capable to distinguish between the different isozymes of the carbonic anhydrase isozymes. As a matter of fact, different conformations such as in the case of CA-II cannot be identified by a metric based on sequence only.

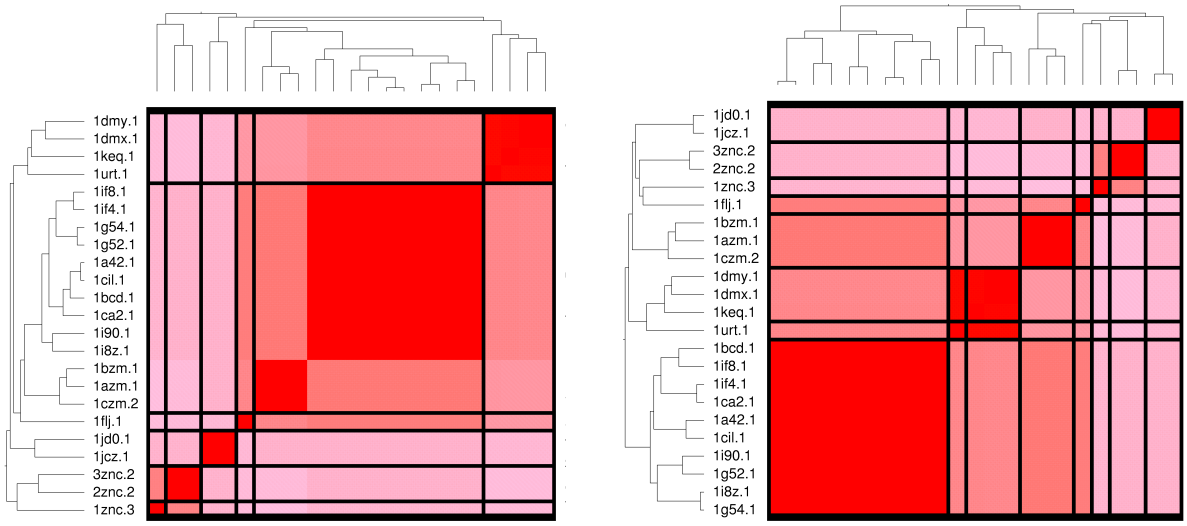


Fig. 5.6 Sequence-based clustering of CAs Clustering of the CA sequences using the clustering algorithm agglo (I) and rb (II).

5.4 Protein kinases

The reversible phosphorylation of proteins is a central control and regulation mechanism in cell growth, signaling and regulation. Kinases catalyze the transfer of a terminal phosphate group from adenosine triphosphate (ATP) to serine, threonine, or tyrosine residues of other proteins. They account for 1.7% of all genes found in the human genome [Manning et al., 2002] and constitute a densely populated protein family. Protein kinases form a large superfamily of proteins. Their catalytic domains are related by sequence and folding similarities (protein kinase fold), nevertheless they exhibit a rich diversity of regulation modes and substrate specificities [Huse and Kuriyan, 2002; Engh and Bossemeyer, 2002; Nolen et al., 2004]. The ATP-binding site is located at the interface between both subdomains (lobes) of the kinase fold. Interference with protein kinase-mediated cell signaling pathways is involved in a variety of diseases including cancer, angiogenesis, neurological diseases, and inflammation [Dancey and Sausville, 2003; Matter, 2001; Saklatvala, 2004]. Consequently, protein kinases have emerged as attractive drug targets for pharmaceutical research. In 2001, the first low-molecular weight ATP-competitive inhibitor (Imatinib, Gleevec®) has been approved as potent agent against chronic myelogenous leukemia (CML). Currently, a vast number of medicinal chemistry projects are followed in the pharmaceutical industry to target the rather homologous ATP-binding site of different kinases with ATP-competitive inhibitors (for recent reviews see [Traxler, 1998; Cohen, 2002; Dancey and Sausville, 2003; Noble et al., 2004]). The elucidation of selectivity-discriminating features present at the binding sites of the different protein kinases is a fundamental prerequisite for developing selective inhibitors, and provides a new taxonomy for the classification of kinases.

Similarity between proteins can be described using different types of information. The most straight-forward approach is the classification of protein kinases based on sequence similarity. Recently, Manning and coworkers presented a classification of 518 non-redundant kinase genes in the human genome (human kinome) [Manning et al., 2002], providing an extensive sequence-based annotation of protein kinases.

A classification focusing on both, sequence and structural information, was performed by Cheek and coworkers [Cheek et al., 2002] for all available sequences of phosphotransfering proteins. The authors clustered over 17000 kinase sequences (also containing kinases acting on non-protein substrates) into 30 distinct families based on sequence similarities. The majority of these sequence families falls into seven general fold groups,

which include the most widespread protein folds (e.g., Rossmann, ribonuclease H, or ferredoxin fold). The protein kinase fold represents one fold subgroup.

A similarity analysis of protein kinases based on folding similarities was presented by Shindyalov and Bourne [Shindyalov and Bourne, 1998]. The authors used a fold comparison algorithm (Combinatorial Extension, (CE)) to align structurally similar regions of different kinase structures. Structural alignments can be calculated for different protein kinases.

There are several methods that use similarities in the folding pattern of protein domains to build up a hierarchical classification such as SCOP (**S**tructural **C**lassification **o**f **P**roteins) [Lo Conte et al., 2002] or CATH (Class(C), Architecture(A), Topology(T) and Homologous superfamily (H)) [Pearl et al., 2000]. Both methods use automatic sequence- and fold-comparison techniques however accomplished by manual annotation. The three main levels ranked, according to ascending similarity in the SCOP hierarchy, are the following: fold (major structural similarity), superfamily (low sequence similarity, but proteins show functional and structural similarity), family (sequence similarity greater than 30%). In comparison, CATH clusters proteins at the four levels mentioned above. These levels reflect the secondary-structure composition and topology.

Besides the classification of proteins based on sequences or secondary structure elements, several approaches focus on 3D structure or the physicochemical properties of their binding sites [Goldsmith-Fischman and Honig, 2003]. These approaches follow the idea that the functional regions of a protein are most important for classification. Recently, a structural classification of protein kinases based on their crystal structures into subfamilies exhibiting similar protein-ligand interactions was performed by Naumann and Matter [Naumann and Matter, 2002]. In an iterative superposition procedure structurally conserved residues were used to obtain a 3D-alignment of the kinase structures. Putative binding properties were mapped using selected probes (N1, O, and DRY, representing hydrogen-bond donor, hydrogen-bond acceptor and hydrophobic interaction properties, respectively) in the GRID force field [Goodford, 1985; Wade and Goodford, 1993]. The 3D-binding site information is encoded in the obtained GRID molecular interaction fields. A principal component analysis (PCA) and consensus PCA [Kastenholz et al., 2000] were used in the analysis of the molecular interaction fields to extract features shared in common by the different kinase binding sites, but in particular to identify those areas that differ among the subfamilies. With the resulting classification (also called 'target family landscapes'), the authors were able to classify the kinases

in the dataset into different subfamilies and rationalize the structural features important for the particular kinase subfamilies. Additionally, 3D-QSAR studies on CDK1 using purine-based inhibitors revealed essential structural elements required for potent inhibition across this family [Naumann and Matter, 2002].

A classification of protein kinases based on small-molecule inhibition data retrieved from the literature was performed by Vieth and coworkers [Vieth et al., 2004]. They used structural similarity and binding profiles of small molecules for the classification of protein kinases and compared these classification results with a sequence-based classification scheme. The authors found that for highly homologous kinases close in sequence space also, on average, the inhibition profiles of small molecule inhibitors correspond closely. However, they were also able to identify cases, where both classifications suggest opposing clustering.

In this paper we present the comparison and classification of protein kinases using the exposed physicochemical properties present in their active sites. This approach does not rely on any given sequence similarities and can be performed in a fully automated fashion.

5.4.1 Initial kinase clustering study

The dataset used by Naumann and Matter [Naumann and Matter, 2002] in their classification study provides a starting point for the cluster analysis with Cavbase. The dataset was extended by other MAP kinases and contained in total 30 kinase cavities[2]. The Cavbase clustering analysis was performed as described in section 5.2. The classification obtained with Cavbase based on the R_1 scoring and allowing for six clusters is shown in Figure 5.7. The clusters (along the diagonal from the lower left to the upper right) consist of cavities from the Fibroblast growth factor receptor kinases, cAMP-dependent protein kinases, tyrosine kinases, CDKs, and MAP kinases Erk2 and p38α (Fig. 5.7). Cavbase is able to clearly distinguish between the different kinase subfamilies. This is an remarkable observation since only structural information about the kinase binding sites was used for the classification with Cavbase. The results are

[2]The dataset comprised the following kinase cavities (see Tab. 5.2 for further details): PKA (1bkx.2, 1atp.2, 1cdk.7, 1bx6.1, 1stc.1, 1ydt.4, 1yds.3, 1fmo.4, 1ydr.3), different Ser/Thr kinases (1hck.1, 1phk.1,1csn.1), Tyr kinases (1ir3.1, 1fgi.1, 1fgi.5, 2src.1, 3lck.1), CDK2 (1ckp.2, 1b38.1, 1fin.5, 1fin.7), and MAP kinases (1gol.1, 1erk.1, 1p38.1, 1pme.1, 3erk.1, 4erk.1, 1bmk.1, 1a9u.1,1bl7.1)

in good agreement with the classifications obtained by methods exploiting information from multiple sources and accomplished by manual intervention (such as CATH or SCOP). Both, the similarity among the MAP kinase cavities from the Erk2 and p38α subfamily is convincingly detected as well as the differences which partitions them from each other (see section 5.4.6 for further details). However, besides an overall correspondence between sequence and cavity space, there are also significant distinctions obvious in both classification schemes. One notable deviation is the differentiation between binding cavities originating from different activation states. Kinases undergo several structural rearrangements while switching between active and inactive states. For example, activation of CDKs requires two steps: binding of cognate cyclin and phosphorylation of a threonine residue (Thr160) residing in the activation loop. Cavbase is able to distinguish kinases of both activation states (active: 1fin.5, 1fin.7 and inactive: 1b38.1, 1ckp.2, 1hck.1). As a matter of fact, such differences cannot be seen in sequence space.

Cavbase is able to detect similarities and differences in the binding sites exhibited by kinases and is capable to provide a reasonable classification of the protein kinases in the considered dataset. As a further step it is interesting to see, how Cavbase performs operating on a significant larger set of protein kinases.

5.4.2 Protein kinase dataset

The kinase structures to be included in the dataset were identified and selected by searching the PDB for structural neighbors with the VAST algorithm [Gibrat et al., 1996] using cAMP-dependent protein kinase 1atp as a query structure. A total of 864 neighboring protein domains were found, originating from 303 different proteins. All non-kinase protein domains and proteins with a resolution beyond 3Å were discarded. In the case of multimeric protein structures, only one molecule per protein entry was considered. The selected dataset comprises 263 kinases from all presently known structural subfamilies (PDB version September 2004).

Table 5.2 lists the kinase entries used in the present Cavbase similarity analysis, together with their corresponding sequence-based group annotation [Manning et al., 2002]

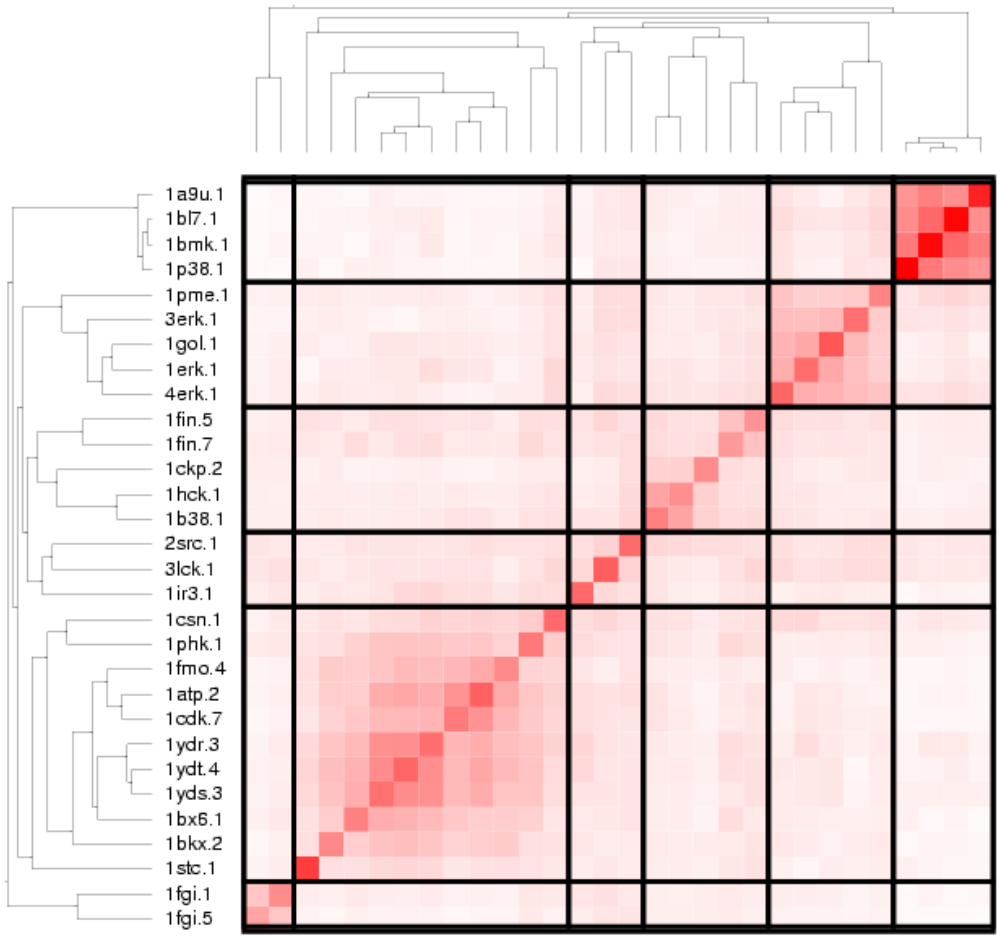


Fig. 5.7 Cavbase clustering for the kinase dataset comprising 30 cavities. The Cavbase clustering solution for 30 kinase cavities is presented. The clustering parameters were selected in a way to achieve separation into 6 distinct clusters, boarders a indicated as solid black lines. The intensity of the reddish coloring indicates the mutual degree of similarity. The displayed clustering solution was obtained using the R_1 scoring function and applying a agglomerative algorithm. Cavbase clusters the 30 kinase structures on the subfamily level.

and the SCOP superfamily classification. Since Cavbase usually detects multiple cavities on the surface of one protein structure, all selected kinase cavities in the dataset were carefully analyzed to confirm that only ATP-binding sites were considered. The largest subfamily are CDK2s (almost 27% of the dataset), followed by protein kinase A, MAP kinases p38 and Erk2, and casein kinases 1 and 2. According to the SCOP classification 52 kinase subfamilies were considered in the dataset.

Tab. 5.2 Kinase structures used in the Cavbase similarity analysis. The table contains the sequence-based family classification [Manning et al., 2002], and the SCOP annotation together with the abbreviation used in this paper for the kinase family, the number of cavities and the PDB accession codes.

SCOP-superfamily[1]	**Sequence group**[2]	**Code**[3]	**No.**	**PDB codes**
Abelsone tyrosine kinase (abl)	TK	cabl	5	1fpu 1iep 1m52 1opj 1opk
Aurora A (aurora-2)	Other	aurora	4	1mq4 1muo 1ol6 1ol7
B-Raf proto-oncogene protein kinase	TKL	braf	1	1uwh
Bruton's tyrosine kinase (Btk)	TK	bruton	1	1k2p
c-jun N-terminal kinase (jnk3s)	CMGC	jnk	5	1jnk 1pmn 1pmq 1pmu 1pmv
C-Kit Kinase Product Complex	TK	ckit	3	1pkg 1t45 1t46
c-src protein tyrosine kinase	TK	csrc	4	1fmk 1ksw 2ptk 2src
Calmodulin-dependent protein kinase	CAMK	camod	1	1a06
cAMP-dependent PK, catalytic subunit	AGC	pkbmut	3	1q24 1q61 1q62
cAMP-dependent PK, catalytic subunit	AGC	camp	26	1apm 1atp 1bkx 1bx6 1cdk 1cmk 1ctp 1fmo 1fot 1j3h 1jbp 1jlu 1l3r 1q8t 1q8u 1q8w 1rdq 1re8 1rej 1rek 1smh 1stc 1ydr 1yds 1ydt 2cpk
Carboxyl-terminal src kinase (csk)	TK	csk	2	1byg 1k9a
Casein kinase-1, CK1	CK1	ck1	5	1cki 1ckj 1csn 1eh4 2csn
Cell cycle checkpoint kinase chk1	CAMK	chk1	4	1ia8 1nvq 1nvr 1nvs
Cell cycle inhibitor p19ink4D	CMGC	cdk6	4	1bi8 1blx 1g3n
Cell division control protein 2 homolog (Pfpk5)	-	cdk2hom	4	1ob3 1v0b 1v0o 1v0p
Choline kinase	-	choline	1	1nw1

(Table continued on next page)

Table continued. 5.2

SCOP-superfamily[1]	Sequence group[2]	Code[3]	No.	PDB codes
Cyclin-dependent PK, CDK2	CMGC	cdk2	71	1aq1 1b38 1b39 1buh 1ckp 1di8 1dm2 1e1v 1e1x 1e9h 1f5q 1fin 1fq1 1fvt 1fvv 1g5s 1gih 1gii 1gij 1gy3 1gz8 1h00 1h01 1h07 1h08 1h0v 1h0w 1h1p 1h1q 1h1r 1h1s 1h24 1h25 1h26 1h27 1h28 1hck 1hcl 1jst 1jvp 1ke5 1ke6 1ke7 1ke8 1ke9 1ogu 1oiq 1oir 1oit 1oku 1ol1 1ol2 1p2a 1p5e 1pf8 1pkd 1pw2 1pxi 1pxj 1pxk 1pxl 1pxm 1pxn 1pxo 1pxp 1qmz 1r78 1urw 1v1k 1vyw 1vyz
Cyclin-dependent PK, CDK5	CMGC	cdk5	1	1h4l
Death-associated protein kinase, Dap	CAMK	dap	4	1ig1 1jkk 1jkl 1jks
EGF receptor tyrosine kinase, Erbb-1	TK	egf	2	1m14 1m17
ephb2 receptor tyrosine kinase	TK	ephb2	1	1jpa
Ephrin A2 (Epha2) Receptor Protein Kinase	TK	epha2	1	1mqb
Fibroblast growth factor receptor 1	TK	fgfr1	4	1agw 1fgi 1fgk 2fgi
Fibroblast growth factor receptor 2	TK	fgfr2	1	1gjo
Fl Cytokine Receptor	TK	flt3	1	1rjb
Focal Adhesion Kinase (Fak)	TK	fak	1	1mp8
G-protein coupled receptor kinase 2, N-terminal domain	AGC	fpcrk2	1	1omw
γ-subunit of glycogen phosphorylase kinase (Phk)	CAMK	phk	3	1phk 1ql6 2phk
Glycogen synthase kinase-3 beta (Gsk3b)	GMGC	gsk3b	11	1gng 1h8f 1i09 1j1b 1j1c 1o9u 1pyx 1q3d 1q3w 1q4l 1uv5
Hemapoetic cell kinase Hck	TK	hck	3	1ad5 1qcf 2hck

(Table continued on next page)

Table continued. 5.2

SCOP-superfamily[1]	Sequence group[2]	Code[3]	No.	PDB codes
Hepatocyte Growth Factor Receptor C-met	TK	cmet	2	1r0p 1r1w
Insulin receptor	TK	insulin	6	1gag 1i44 1ir3 1irk 1p14 1rqq
Insulin-like growth factor 1 receptor	TK	inslik	4	1jqh 1k3a 1m7n 1p4o
Lymphocyte kinase (lck)	TK	lck	5	1qpc 1qpd 1qpe 1qpj 3lck
MAP kinase activated protein kinase 2, mapkap2	CAMK	mapkap2	3	1kwp 1nxk 1ny3
MAP kinase Erk2	CMGC	erk2	6	1erk 1gol 1pme 2erk 3erk 4erk
MAP kinase p38	CMGC	p38a	18	1a9u 1bl6 1bl7 1bmk 1di9 1kv1 1kv2 1lew 1lez 1m7q 1ouk 1ouy 1ove 1oz1 1p38 1r39 1r3c 1wfc
MAP kinase p38-gamma	CMGC	p38g	1	1cm8
Musk tyrosine kinase	TK	musk	1	1luf
Mycobacterial protein kinase PknB, catalytic domain	Other	pknb	2	1mru 1o6y
P3-phosphoinositide dependent protein kinase-1 Pdk1	AGC	pdk1	6	1h1w 1uu3 1uu7 1uu8 1uu9 1uvr
pak1	STE	pak1	1	1f3m
Pkb kinase	AGC	pkb	7	1gzk 1gzn 1gzo 1mrv 1mry 1o6k 1o6l
Protein kinase CK2, alpha subunit	Other	ck2	14	1daw 1day 1f0q 1j91 1jam 1lp4 1lpu 1lr4 1m2p 1m2q 1m2r 1na7 1om1 1pjk
Sky1p	Other	sky1p	5	1how 1q8y 1q8z 1q97 1q99
Tie2 kinase	TK	tie2	1	1fvr
Titin, kinase domain	CAMK	titin	1	1tki
Type I Tgf-beta Receptor	TKL	tgf1	2	1b6c 1ias
Twitchin, kinase domain	CAMK	twitchin	1	1kob

(Table continued on next page)

Table continued. 5.2

SCOP-superfamily[1]	Sequence group[2]	Code[3]	No.	PDB codes
Vascular endothelial growth factor receptor 2 (kdr)	TK	kdr	1	1vr2

[1] Superfamily for given catalytic kinase domain extracted from the SCOP database release 1.65. If no SCOP superfamily is given for a particular protein kinase, the information is adapted from the PDB database.

[2] Sequence-based classification data taken from [Manning et al., 2002] (www.kinase.com/human/kinome). Abbreviations for sequence groups: AGC containing PKA, PKG, PKC families; CAMK calcium/calmodulin-dependent protein kinase; CK1 casein kinase 1; CMGC containing CDK, MAPK, GSK3, CLK families; STE homologs of yeast sterile 7, sterile 11, sterile 20 kinases; TK tyrosine kinase; TKL tyrosine kinase like.

[3] Abbreviations for kinase family used in this work, e.g. in Table 5.3

5.4.3 Focussing on the ATP-binding site

For clarity, the ATP-binding site will be categorized into five regions following the pharmacophore model introduced by Traxler and using the ATP binding mode as reference: the adenine region, the ribose pocket, the hydrophobic region I and hydrophobic region II (also referred to as the specificity surface), and the phosphate binding groove [Traxler, 1998; Cherry and Williams, 2004] (Figure 5.9, (II)). Crystal structure analysis reveals that ATP leaves both hydrophobic regions unoccupied. Consequently, inhibitors addressing these areas can establish selectivity determining features towards different kinases.

In addition to the five pharmacophoric regions, a kinase ATP-binding site as extracted by Cavbase comprises also parts of the catalytic and activation loop. An ATP binding cavity extracted from a cAMP dependent protein kinases (PKA, PDB code 1atp) is shown in Figure 5.8. Having a kinase classification in mind that captures the most important features for kinase selectivity and subfamily characterization, we focus our analysis on residues next to the ATP-binding site. Accordingly, the amino acids involved in the hinge hydrogen-bonding network (defined as residues structurally corresponding to the peptide backbone carbonyl (Glu121) and the peptide amide nitrogen (Val123) in PKA (1atp)) are identified for each kinase (Table 5.3). In the mutual comparison of two kinase cavities, only those pseudocenters that coincide with a predefined sphere around the hinge region are considered. In three analysis scenarios the radius of this extraction sphere was set to 9.0Å, 12.0Å or 16.0Å (Figure 5.9). The radius was adjusted such that residues forming (i) the adenine and both hydrophobic regions, (ii) in addition to the latters the sugar pocket, and (iii) finally further more the phosphate binding groove are included into the similarity analysis (9.0Å, 12.0Å and 16.0Å, respectively). This approach is more strongly focused on the essential discriminating features in the ATP-binding site. In this way, the classification is less dependent on the different conformations exhibited by the activation loop. This loop is known to be able to undergo drastical movements, and different conformations experienced by the activation loop might dominate the results of our classification analysis.

kinase family	peptide CO	peptide NH	kinase family	peptide CO	peptide NH
camp	Glu121	Val123	camp2	Asp165	Ile167
ck1	Asp86	Leu88	cyclin	Glu81	Leu83
cdk2	Glu81	Leu83	erk2	Asp104	Met106
erk2mut	His106	Met108	p38	His107	Met109
lck	Glu317	Met319	insulin	Glu1077	Met1079
inslik	Glu1080	Met1082	inslik2	Glu1050	Met1052
inslik3	Glu1077	Met1079	src	Glu339	Met341
gsk3b	Glu133	Val135	abl	Glu316	Met318
abl2	Glu335	Met337	dap	Glu94	Leu95
mapkap2	Glu139	Leu141	hck	Glu339	Met341
fgwr	Glu562	Ala564	jnk3	Glu147	Met149
cdk6	Glu99	Val101	src	Glu339	Met341
phk	Asp104	Met106	pkb	Glu/Asp230	Ala232
pkbmut	Glu121	Ala123	p38γ	Pro110	Met112
sky1p	Glu247	Leu249	ck2	Glu114	Val116
pdk1	Ser160	Ala162	auroraA	Glu211	Ala213
aurrel	Glu211	Ala213	bruton	Glu475	Met477
csk	Glu267	Met269	choline	Glu152	Ile154
ckit	Glu671	Met673	ephb2	Glu708	Met710
fpcrk2	Asp267	Met274	cmet	Pro1158	Met1160
titin	Glu68	Ile332	twitchin	N/A	Asn202
tgf1	Asp121	His283	amipt	Ser91	Ala93
kdr	Glu917	Cys919	cdk2hom	Glu80	Leu82

Tab. 5.3 Residues in selected protein kinases involved in the hinge hydrogen bonding network with ATP. Multiple entries for a subfamily are listed, if different residues of one protein subfamily are involved in the hydrogen bonding.

5.4.4 Sequence based clustering and SCOP classification

The obtained Cavbase classification is compared to a sequence-based and SCOP- based classification to assess the quality of our clustering results. For the sequence-based clustering all kinase sequences of the catalytic domains were extracted using Relibase. Subsequently, they were mutually aligned using BLAST [Altschul et al., 1997] in standard settings, with the exception of an increased E-value (set to 100). Additionally, the kinase sequences were compared using FASTA [Pearson and Lipman, 1988]. The BLAST bit score and the sequence identity values obtained from FASTA were used to build up the similarity matrix, which was subjected to the same clustering procedures applied to the Cavbase output.

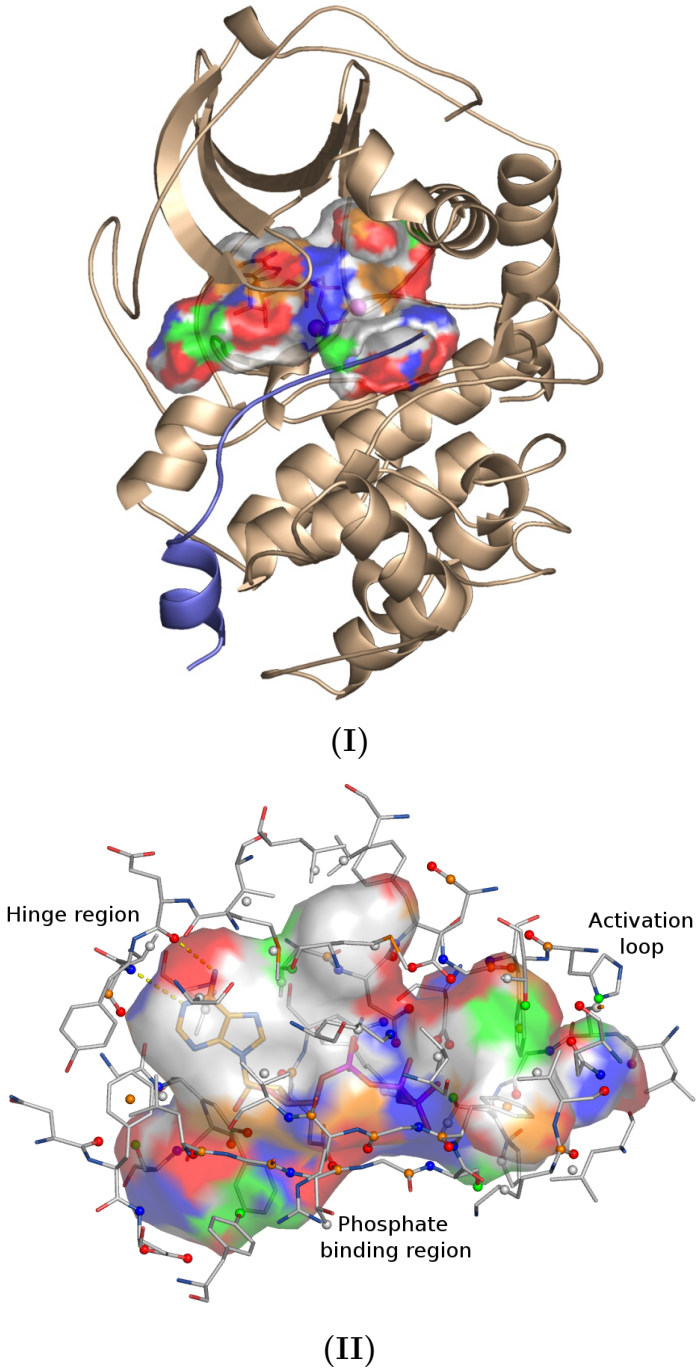


Fig. 5.8 Typical protein kinase fold of a cAMP-dependent protein kinase (PDB code 1atp) and the ATP-binding site as detected by Cavbase. The bilobal protein kinase fold of the predominantly α-helical N-terminal lobe and the C-terminal lobe is shown in (I). The ATP-binding site is located at the interface. The surface of the ATP cavity as detected by Cavbase is displayed, the bound protein-like substrate is colored in blue. The Cavbase cavity comprises the functionally important regions of the binding site (II): the adenine-binding pockets, the hydrophobic pocket I and II, the sugar and phosphate binding region, the DFG motif, the glycine flap and parts of the activation and catalytic loop (the ATP cavity is rotated around 40^o compared to (I)). The ATP cavity is rather narrow and deeply buried and has next to the adenine moiety a predominantly hydrophobic character (white colored surface).

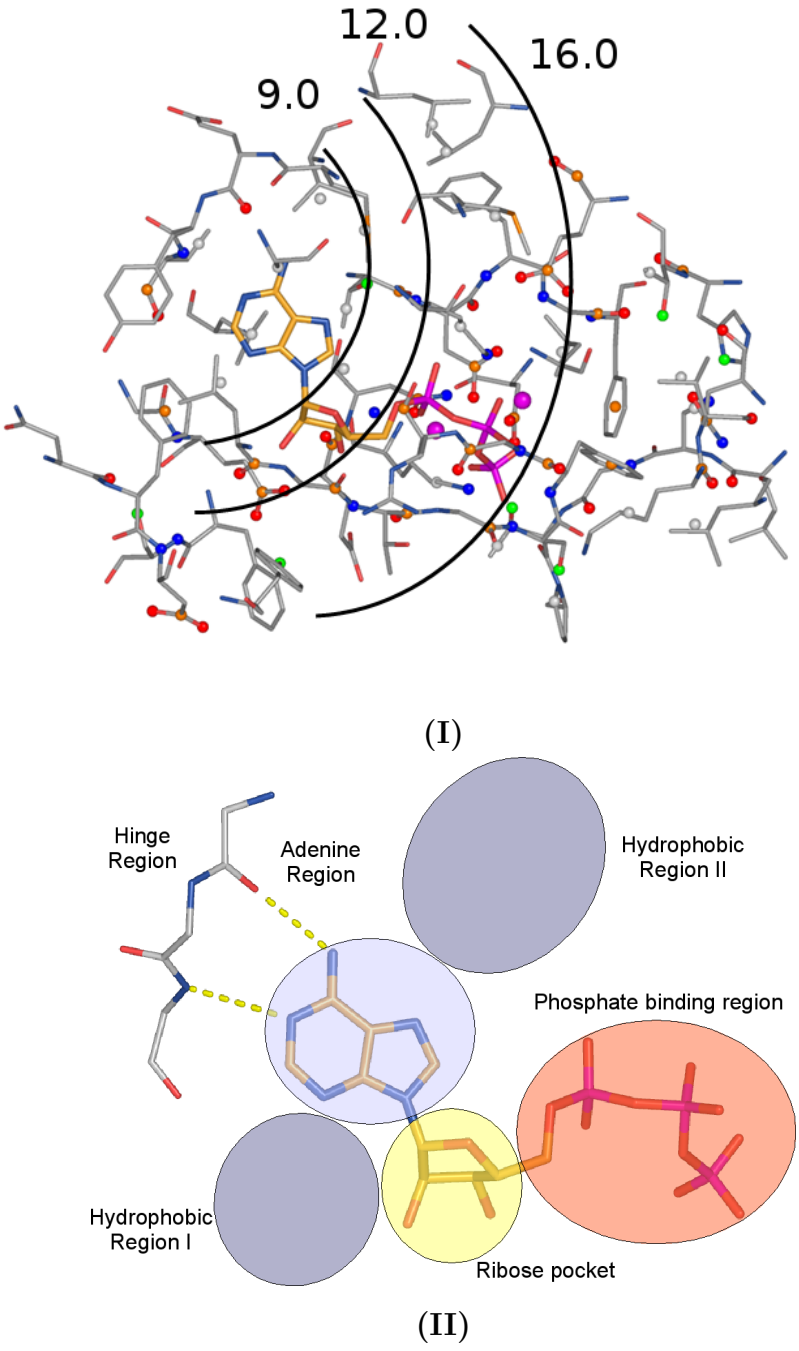


Fig. 5.9 Focusing on regions of the kinase binding site relevant for inhibitor binding. To concentrate on the adenine pocket and to better reflect the selectivity determining features with respect to small molecule binding, the cavities of the kinases are limited to the ATP-pocket. Pseudocenters are excluded from the cavity comparison, if they are more than 9.0Å, 12.0Å, or 16.0Å distant from the hinge region (I). (II) displays the kinase pharmacophore for an EGF-kinase adapted from Traxler et al. [Traxler, 1998] categorizing the ATP-binding site into five structural regions which have a distinct chemical character. In all the figures this view will be used unless noted otherwise.

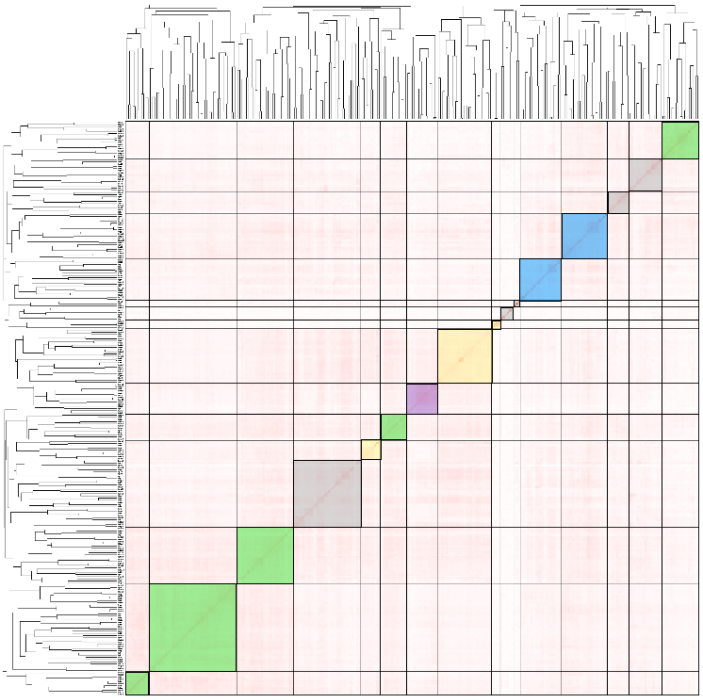


Fig. 5.10 Clustering of the Cavbase similarity scoring of 263 binding cavities. The clustering parameters were setup in a way to achieve separation into 16 distinct clusters. The displayed clustering solution was obtained using the R_1 scoring function and applying a partitional clustering strategy. The sequence-based classification given by Manning and coworkers [Manning et al., 2002] arranges the protein kinases (human kinome) into eight sequence groups. The latter classification scheme is mapped onto the Cavbase results by applying the following color coding: AGC (yellow), CMGC (green), TK (blue), and Other (violet). Cavities corresponding to these sequence groups are found in clusters consisting of cavities from one particular group alone. Clusters, which contain cavities from multiple sequence groups are colored in grey (e.g., clusters containing cavities corresponding from the sequence groups STE, CK1, and TKL). In general Cavbase groups cavities from one sequence class together. Allowing for higher number of clusters, the Cavbase clustering distinguishes between the different kinase examples already on the subfamily level.

5.4.5 Overall analysis of the clustering results

As in the case of the smaller kinase dataset (see section 5.4.1), the class of serine/threonine and tyrosine protein kinases are separated into different clusters by the Cavbase analysis. The results of a Cavbase clustering analysis based on the R_1 scoring and allowing for 16 clusters are shown in Figure 5.10. The sequence-based group annotation from Manning et al. [Manning et al., 2002] is mapped onto the Cavbase classification. Facing both classification schemes it is indicated that a high sequence similarity between two protein kinases generally parallels high structural similarity of their binding sites. Binding cavities from kinases that show high sequence identity tend to be similarly clustered by the Cavbase approach (Table 5.4). This observation - although it might sound trivial - just underscores the fact that Cavbase is able to detect existing similarities across related binding sites. Highly related proteins also exhibit similar binding sites. Even for the largest kinase subfamily in the dataset (CDK2), the similarity in the active sites is detected and all CDK2s are separated from the other cavities in the dataset. Furthermore a clear distinction into clusters consisting of CDK2s exhibiting active or inactive conformations is achieved. This observation becomes even more striking, since 71 kinase cavities from this protein subfamily were used in the classification analysis.

The Cavbase classification differs from classification schemes based on sequence information in the way how relationships between different kinase subfamilies are established. Cavbase detects kinases that exhibit different degrees of similarity in sequence and Cavbase space. For example, in section 5.4.8 the binding of Gleevec to different kinases is rationalized using similarities in their binding sites with Cavbase. Additionally, section 5.4.9 provides several examples for pairs of kinases, which behave significantly different in sequence and Cavbase space.

Interestingly enough, the classification between different kinases into subfamilies and distinct activation states is already achieved when evaluating the reduced kinase subcavities limited to the adenine region (residues coinciding with a 9.0Å sphere around the hinge region). Kinases from one subfamily show already significant similarities next to the adenine binding region to other kinases cavities of the same subfamily. In consequence, all cavities from such a subfamily are found in one single cluster (Table 5.4). It supports the hypothesis that even in the relatively homologous ATP-binding site,

a structural distinction between kinases from different subfamilies is possible using Cavbase.

In the following sections, the use of Cavbase for the analysis and visualization of active site similarities and differences is shown for two prominent subfamilies: the MAP kinases 5.4.6 and c-Abl kinases 5.4.7. The results of the presented clustering analysis were obtained using the dataset of 263 kinases (except of the clustering of the MAP kinase sub-dataset) and the cavities limited to the ATP-pocket (16Å) (except otherwise noted) as input.

Tab. 5.4 A selection of kinase sequence subfamilies that cluster according to 3D-similarity in one single cluster consisting only of cavities from that subfamily. The similarities are also detected, when using the cavities limited to the adenine region.

Kinase SCOP superfamily
3-phosphoinositide dependent protein kinase-1 Pdk1
Cell cycle checkpoint kinase chk1
c-jun N-terminal kinase (jnk3s)
EGF receptor tyrosine kinase, Erbb-1
Glycogen synthase kinase-3 beta (Gsk3b)
Lymphocyte kinase (lck)
MAP kinase Erk2
MAP kinase p38

5.4.6 MAP kinases

Mitogen-activated protein kinases (MAPKs) are involved in signal-transduction cascades that control complex processes such as cell proliferation, differentiation, and apoptosis. Based on their signature activation sequences, MAPKs can be categorized into at least three broad subfamilies: MAP kinase p38, extracellular-signal-regulated kinases (ERKs), c-jun amino-terminal kinases (JNKs) [Schramek, 2002; Johnson and Lapadat, 2002]. Especially the MAP kinase p38α subfamily is of pharmaceutical interest since these kinases are involved in the production of proinflammantory cytokines [Kumar et al., 2003]. Thus, selective inhibition of the p38α isoform provides a possible therapeutic intervention for a wide range of inflammatory diseases.

The dataset used for the similarity analysis (263 kinase cavities) contains a subset of 30 MAP kinases originating from the p38, Erk2, and JNK3 subfamilies as listed in Table 5.5. Most MAP kinase structures were determined in the unphosphorylated, inactive state. The Cavbase analysis based on all 263 cavities considering residues coinciding within a 16.0Å sphere around the hinge region and a number of 48 clusters reveals that Cavbase separates the 30 different MAP kinases on the subfamily level in four groups: p38α, p38γ, Erk2, and JNK3. In detail, the p38α cavities are found in four clusters consisting solely of p38α cavities, the Erk2 cavities are found in one cluster populated only with Erk2 cavities and the JNK3 cavities are found in two cluster grouped together with casein kinase cavities. To further investigate the cross-relationships between the MAP kinase subfamilies, a clustering analysis was performed using only the 28 MAP kinase structures as input[3]. The classification results for the Cavbase classification acting on this subset of MAP kinases are shown in detail in Figure 5.11 using the Cluto's mountain visualization option. The number of clusters was set to ten. According to the Cavbase classification, the cavities are also grouped on the subfamily level (p38, Erk2, JNK3). The cluster analysis allows one to navigate through the MAP kinase family, and to visualize cross relationships between them (see legend Figure 5.11 for further details).

An interesting example with respect to the analysis of functional cross-relationships between protein families is the penta mutant of a MAP Erk2 kinase (PDB code 1pme [Fox et al., 1998]). This mutated Erk2 kinase was constructed to mimic the binding

[3]The p38α cavities from PDB entries 1r39 and 1r3c were not included in this cluster analysis. These cavities were added to the big kinase dataset after this analysis was performed.

pocket of MAP p38α based on the Erk2 skeleton. The comparison of this Erk2 kinase with respect to the other MAP kinases reveals the transient character of this penta mutant. Cavbase detects similarities either to Erk2 and p38α structures and clusters this cavity half way in between both classes (Figure 5.12).

The next two examples show the usage of Cavbase in the graphical analysis of similarities and dissimilarities in the binding sites of two proteins. Cavbase easily superimposes two cavities based on similar areas in their active sites and is able to facilitate the rationalizing of features important in discrimination between both binding sites.

In the Cavbase classification, cavities from the Erk2 and p38α are separated as expected. What are the differences between both subfamilies? Inhibition profiles of small molecules reveal that there are several compound classes showing a selective inhibition of p38α, but they do not inhibit Erk2. One example of such binders are pyridinylimidazole type inhibitors introduced by SmithKline Beecham, e.g. compound SB203580 ((**13**), [Lee et al., 1994; Cuenda et al., 1995; Wang et al., 1998]). This class shows high selectivity for p38α compared to other MAP kinases. These binding properties can be explained by identifying the gatekeeper residue and analyzing its influence on the size of the hydrophobic pocket II (Figure 5.12) [Wang et al., 1998; Lee et al., 1999]. In p38α, the size of this pocket is controlled by a small residue (threonine), whereas other MAP kinases exhibit a larger residue at this position (e.g. it is a glutamine in Erk2) [Gum et al., 1998; Lee et al., 1999]. The Cavbase analysis shows that there are similarities around the hinge hydrogen-bonding region and the adenine binding area, but both cavities exhibit no similarity in the hydrophobic region II near the gatekeeper residue (Figure 5.12).

A further example of an experimentally observed inhibitor specificity for p38α over other MAP kinases could also be rationalized with Cavbase. The quinazolinone- (**16**) and dihydropyridopyrimidinone-based inhibitors (**15**) show selective inhibition of p38α [Fitzgerald et al., 2003] and several crystal structures of complexes with these inhibitors have been determind. Besides the influence of the size deviating gatekeeper residues present in Erk2 and p38α, a peptide flip in the protein backbone between Met109 and Gly110 is proposed to explain the unexpected specificity exhibited by dihydropyridopyrimidinone-based inhibitors (**16**) [Fitzgerald et al., 2003]. These inhibitors possesses a unique hydrogen-bond acceptor functional group (keto group) at the position where most kinase inhbitors exhibit a donor function (e.g. a pyrimidine nitrogen), which is able to form a hydrogen bond to the hinge region. This deviating

pattern of donor/acceptor properties oriented towards the hinge region is specific for this compound class compared to other kinase inhibitors and particularly addresses the flipped peptide backbone orientation in the p38α kinase family, which is selectively induced in the p38 kinase upon inhibitor binding. A glycine residue next to Met109 in MAP p38α facilitates this flip, whereas larger residues present at this position in other MAP kinases hamper the flipped orientation. Cavbase is able to capture this deviating H-bonding pattern next to the Met109 backbone region that results from the flipped peptide orientation in p38 compared to other MAP kinases. The corresponding regions in the active site of two MAP p38α kinases in complex with a pyrido-pyrimidine and a dihydroquinazolinone ligand are shown in Figure 5.12.

Tab. 5.5 MAP kinases used in the classification analysis.

PDB code	MAP subfamily	complex structure	activity state	Reference
1erk	Erk2	apo-structure	inactive	[Zhang et al., 1994]
1gol	Erk2	binary MgATP	inactive	[Robinson et al., 1996]
1pme	Erk2	pyridinyl-imidazole	inactive	[Fox et al., 1998]
2erk	Erk2	apo-structure	active	[Canagarajah et al., 1997]
3erk	Erk2	pyrimidinyl-imidazole	inactive	[Wang et al., 1998]
4erk	Erk2	olomoucine	inactive	[Wang et al., 1998]
1a9u	p38α	pyridyl-imidazole	inactive	[Wang et al., 1998]
1bl6	p38α	pyridyl-imidazole	inactive	[Wang et al., 1998]
1bl7	p38α	pyrimidinyl-imidazole	inactive	[Wang et al., 1998]
1bmk	p38α	pyrimidinyl-imidazole	inactive	[Wang et al., 1998]
1di9	p38α	4-anilinoquinazoline	inactive	[Shewchuk et al., 2000]
1kv1	p38α	diaryl-urea	inactive	[Pargellis et al., 2002]
1kv2	p38α	diaryl-urea	inactive	[Pargellis et al., 2002]
1lew	p38α	with substrate	inactive	[Chang et al., 2002]
1lez	p38α	with substrate	inactive	[Chang et al., 2002]
1m7q	p38α	dihydro-quinazolinone	inactive	[Stelmach et al., 2003]
1ouk	p38α	pyrimidinyl-imidazole	inactive	[Fitzgerald et al., 2003]
1ouy	p38α	dihydro-pyridopyrimidinone	inactive	[Fitzgerald et al., 2003]
1ove	p38α	dihydro-pyridopyrimidinone	inactive	[Fitzgerald et al., 2003]
1oz1	p38α	4-azaindoles	inactive	[Trejo et al., 2003]
1p38	p38α	apo-structure	inactive	[Wang et al., 1997]
1r39	p38α	apo-structure	inactive	[Patel et al., 2004]
1r3c	p38α	apo-structure	inactive	[Patel et al., 2004]
1wfc	p38α	apo-structure	inactive	[Wilson et al., 1996]
1cm8	p38γ	binary MgANP	active	[Bellon et al., 1999]
1jnk	JNK3	MgANP	inactive	[Xie et al., 1998]
1pmn	JNK3	pyrimidinyl-imidazole	inactive	[Scapin et al., 2003]
1pmq	JNK3	pyrimidinyl-imidazole	inactive	[Scapin et al., 2003]
1pmu	JNK3	phenanthroline	inactive	[Scapin et al., 2003]
1pmv	JNK3	dihydroanthra-pyrazol	inactive	[Scapin et al., 2003]

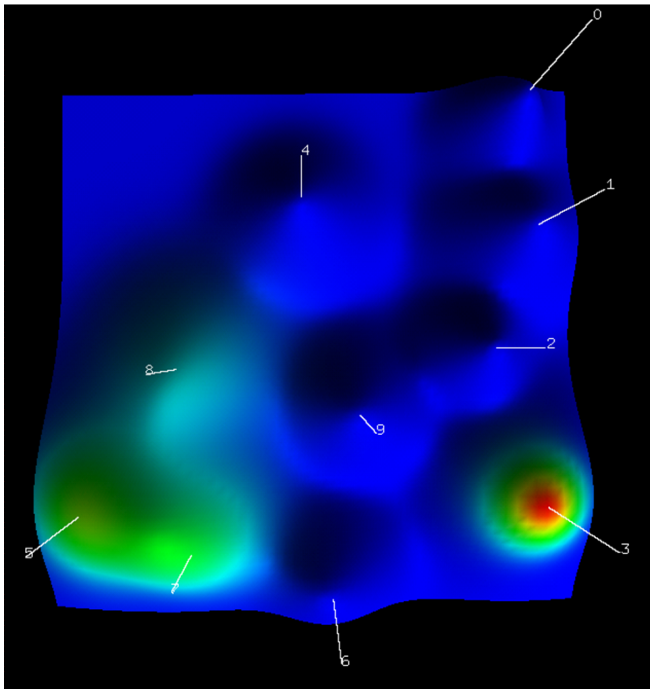


Fig. 5.11 Cavbase classification of MAP kinases. The classification results for all MAP-kinases using the mountain visualization option in Cluto [Karypis, 2002] are displayed. In this 3D plot, the relative similarity of the clusters together with their size, internal similarity, and internal deviation are displayed. Each cluster is represented as a peak in the plane, where the height of a cluster is proportional to the internal homogeneity of the cluster and the volume of a cluster correlates with the number of cluster members. The color at the peak of a cluster is commensurate to the internal deviation (red color refers to low deviation; blue refers to high deviation; [Karypis, 2002]. The MAP p38α kinases are found in the clusters labelled with 4, 5, 6, 7 and 8, cluster 3 is entirely composed by Erk2 entries, the JNKs are found in cluster 1 and 4. The Erk2-pentamutant 1pme mimicking the p38α is found together with an Erk2 structure (PDB code 3erk) between the p38 (cluster 4,6,8) and Erk2 (cluster 3) clusters, reflecting the highest similarity of an Erk2 kinase with a p38α kinase (cluster 9). The pentamutant 1pme and the 3erk structure have both a p38α-specific pyridinyl-imidazole-based inhibitor bound.

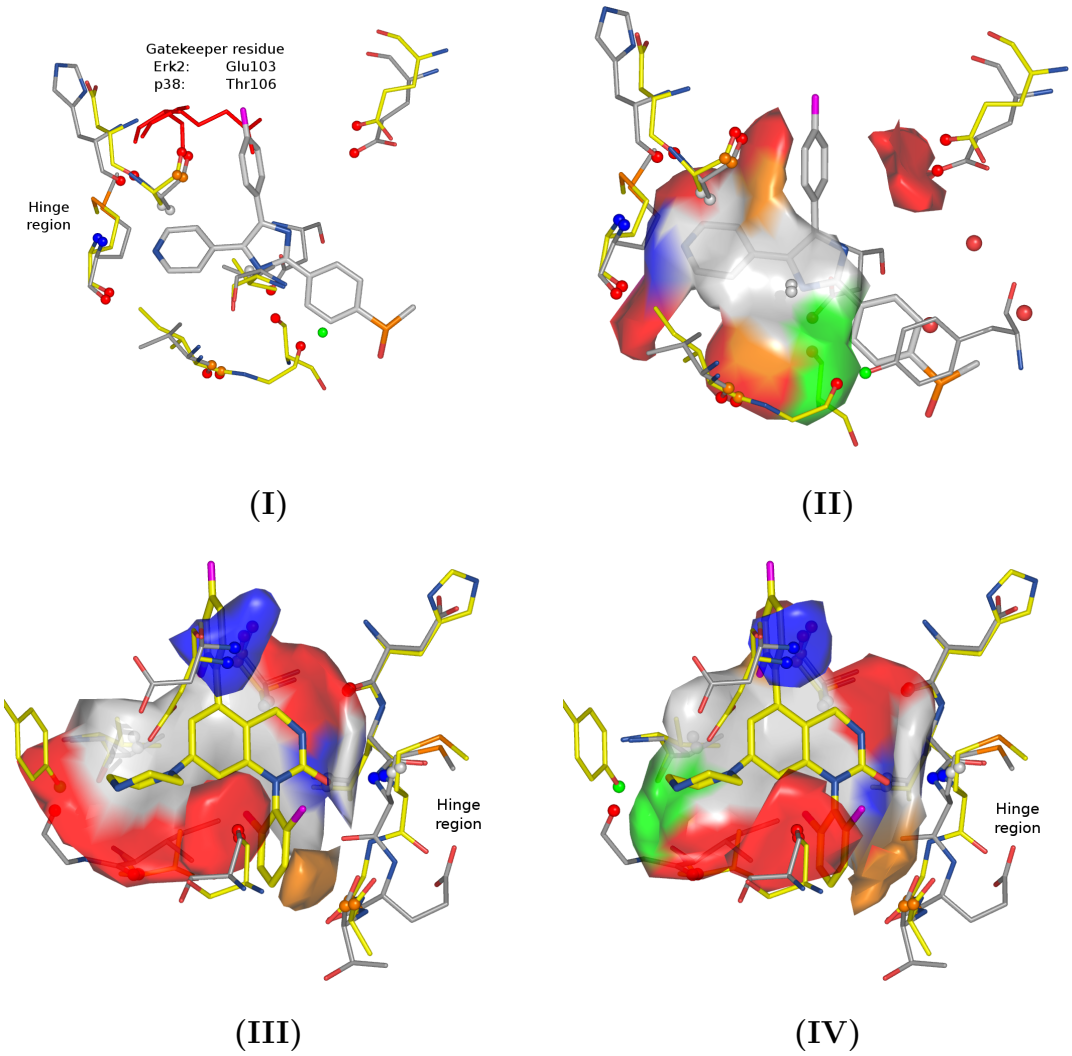
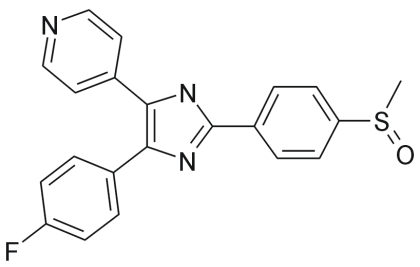
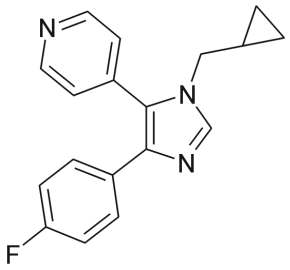


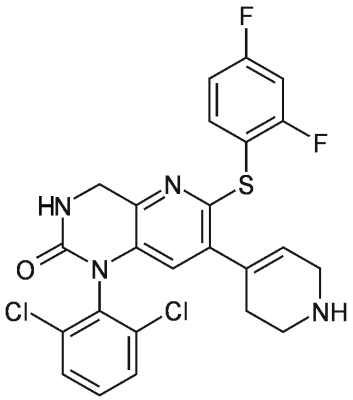
Fig. 5.12 Selectivity determining features in the binding sites of MAP kinases of the p38α and Erk2 subfamliy. Pyrido-imidazole-based inhibitors are selective for p38α. In (I) the possible clash of the gatekeeper residue in Erk2 (Glu103 in Erk2, Thr106 in p38α, both residues colored in red) with the fluorophenyl ring of the SB inhibitor is shown. (II) displays the corresponding amino acids, pseudocenters and one similar cavity surface (only shown from 1a9u) in the binding site of a p38α kinase (PDB code 1a9u; carbon atoms colored in white) and an Erk2 kinase (PDB code 1erk; carbon atoms colored in yellow). The pyridinyl-imidazole inhibitor SB216995 (**14**, see 5.13) forms a hydrogen bond with the pyridine N1 to the Met109 backbone amide. The fluorophenyl ring binds in the hydrophobic pocket near the sidechains of Thr106 (gatekeeper). Cavbase detects no similarity between both proteins in that region. In (III) and (IV) the use of Cavbase for the analysis of subtle differences either from an Erk2 or p38α is shown. Dihydroquinazolinone-based inhibitors (e.g., (**16**)) show a unique hydrogen-bonding pattern to the hinge. The carbonyl group acts as hydrogen-bond acceptor, most kinase inhibitors show a hydrogen-bond donor at this position. To avoid repulsion, the protein backbone of a p38α kinase performs a peptide flip to accommodate this inhibitor. This flip is facilitated by the presence of a glycine residue next to Met109, larger residues would hamper the occurrence of a ligand-induced peptide flip [Fitzgerald et al., 2003]. In (III) and (IV) the similar areas in the binding sites of an Erk2 (PDB code 1erk, carbon atoms colored in grey) and p38α kinase are shown (PDB code 1m7q, carbon atoms colored in yellow). In each figure only one cavity surface from either Erk2 or p38 is shown respectively. The lack of similarity next to the flipped peptide bond can be easily detected.



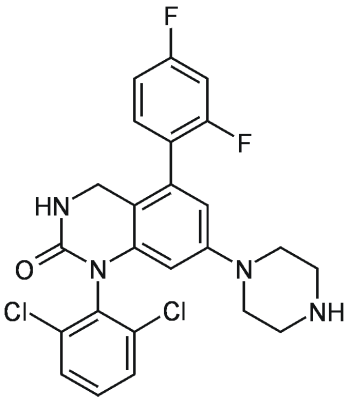
13



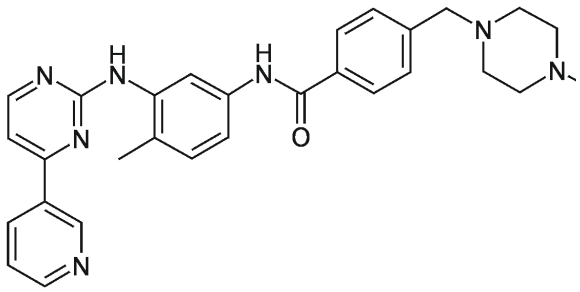
14



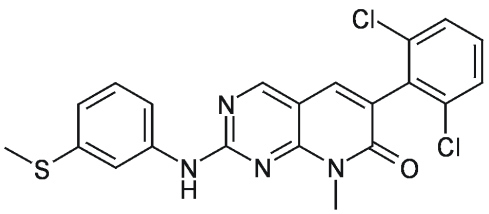
15



16



17



18

Fig. 5.13 Low molecular weight ATP-competitive kinase inhibitors with known binding modes. The SB compounds SB203580 (**13**) and SB216995 (**14**) belong to the pyridinyl-imdidazole-based class of inhibitors. They are selective inhibitors of MAP kinase p38α. Equally, the dihydroquinazolinone-based inhibitors (**15**) and (**16**) are selective inhibitors of MAP kinase p38α. STI-571 (**17**) (Imatinib, Gleevec) is a potent inhibitor of c-Abl kinase, platelet-derived growth factor receptor (PDGF-R) kinase α and β, and the c-Kit kinase. The pyrido[2,3-d]pyrimidine compound PD173955 (**18**) is a classical hinge-binder and has a greater potency towards c-Abl, but is a less selective inhibitor.

5.4.7 c-Abl tyrosine kinases

The cellular form of the Abelson leukemia virus tyrosine kinase (c-Abl) belongs to the class of non-receptor tyrosine kinases. Patients suffering from chronic myelogenous leukemia (CML) often possess a reciprocal translocation between chromosomes 9 and 22 that leads to the formation of a chimeric gene (Bcr-Abl) on the so-called Philadelphia chromosome [Rowley, 1973; Druker, 2003]. The resulting fusion protein (Bcr-Abl) has an intact Abl kinase domain but lacks the internal control mechanism that keeps c-Abl in an inactive form [Druker, 2003]. Increased tyrosine kinase activity of the Bcr-Abl protein leads to CML [Lugo et al., 1990]. 2-Phenylaminopyrimidine-based inhibitors like STI-571 (Gleevec, approved for therapy against CML) have a high affinity towards Abl kinase [Capdeville et al., 2002].

Five c-Abl structures are contained in the present dataset (263 kinases), two structures with pyrido-[2,3.d]pyrimidine-based inhibitors (PDB code 1m52 and 1opk) and three structures with 2-phenylaminopyrimidine-based inhibitors (PDB code 1iep, 1fpu, 1opj). Cavbase is able to detect their similarity in the ATP-binding sites, and all five c-Abl binding sites are clustered together (Fig. 5.14). Additionally, an example for stem-cell factor receptor c-Kit (a receptor tyrosine protein kinase) (PDB code 1t46) with bound STI-571 (**17**) is found in this cluster. Cavbase further groups the c-Abl cavities with

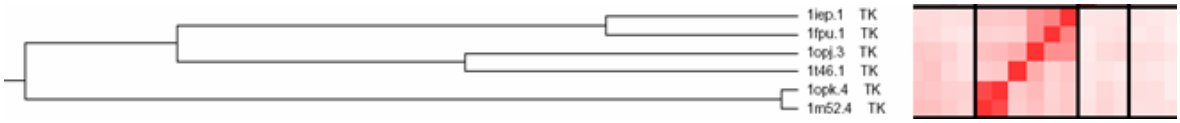


Fig. 5.14 Cluster with c-Abl cavities. Cavbase detects the similarity between the five c-Abl cavities in the present dataset of 263 kinase structures and groups them into one cluster (number of output clusters was set to 38). Furthermore, Cavbase distinguishes between the c-Abl structure in an active conformation (1m52 and 1opk) and those in the inactive conformation (1fpu, 1iep and 1opj). Interestingly, a c-Kit kinase cavity (1t46) with STI-571 bound is found in the same cluster (see chapter 5.4.8).

the pyrido-[2,3.d]pyrimidine-based inhibitors (PD173955 (**18**) and PD166326) (Figure 5.13) and the other c-Abl and the c-Kit cavity with STI-571 or a STI-571-analog bound together. This separation reflects the different activation states of the c-Abl kinases. STI-571 traps the c-Abl kinase in an inactive conformation [Nagar et al., 2002]. It is believed that STI-571 cannot bind to the active conformation of c-Abl because of a steric clash with the DFG motif (as discussed below). The smaller PD173955 is more potent against c-Abl than STI-571. This observation suggests that PD173955 is able

to bind either to the inactive conformation as well as the active conformation of c-Abl, thus recognizing multiple conformational states of this kinase. However, PD173955 is a less selective inhibitor. Structurally, the kinase domain in both complexes with either STI-571 or PD173955 [Nagar et al., 2002] is virtually identical, but there are large differences in the conformations of the activation loop and also rearrangements of the DFG motif. c-Abl kinase cavities in the activated state accommodating pyrido-[2,3.d]pyrimidine-based inhibitors show a substantial degree of similarity in the active site (Figure 5.15-I, Table 5.6). The protein environment and the bound inhibitors are nearly perfectly superimposable. In contrast, the comparison of the active and the inactive conformations of c-Abl kinases reveals similarity only next to the hinge region and the hydrophobic region II. However there are significant structural differences near the DFG motif, which adopts a very different conformation in the active and inactive states. In the active state, the aspartate of this motif points into the cavity towards the magnesium ions (DFG-in conformation), whereas it orients off from the ATP-binding site in the inactive state (DFG-out). Figure 5.15-II shows a superposition of both cavities. Cavbase detects the areas, which are similar in both conformations and detects these regions, where differences occur. The similar areas in the ATP-binding site are located next to the hinge region and the hydrophobic pocket. The difference in the conformation of the DFG motifs preventing STI-571 from binding to the active conformation (DFG-in) can be easily recognized.

Tab. 5.6 As equivalent detected areas in the binding site of c-Abl and c-Kit.

c-Abl (1opj)			**c-Kit** (1t46)		
pseudocenter type	equivalent amino acid [a]		pseudocenter type	equivalent amino acid[a]	
pi	A 288	p	pi	A 621	p
acceptor	A 288	p	acceptor	A 621	p
pi	V 289	p	pi	V 622	p
donor	K 290	p	donor	K 623	p
acceptor	E 305	s	acceptor	E 640	s
pi	E 305	s	pi	E 640	s
donor	V 318	p	donor	V 654	p
acceptor	I 332	p	acceptor	V 668	p
donor	T 334	p	donor	T 670	p
don_acc	T 334	s	don_acc	T 670	s
acceptor	E 335	p	acceptor	E 671	p
aromatic	F 336	s	aromatic	Y 672	s
donor	M 337	p	donor	C 673	p
acceptor	M 337	p	acceptor	C 673	p
pi	G 340	p	pi	G 676	p
donor	G 340	p	donor	G 676	p
don_acc	H 380	s	don_acc	H 790	s
pi	V 398	p	pi	I 808	p
acceptor	V 398	p	acceptor	I 808	p
pi	A 399	p	pi	C 809	p
donor	D 400	p	donor	D 810	p
acceptor	D 400	p	acceptor	D 810	p
pi	D 400	s	pi	D 810	s
aliphatic	L 267	s	aliphatic	L 595	s
aliphatic	V 275	s	aliphatic	V 603	s
aliphatic	A 288	s	aliphatic	A 621	s
aliphatic	M 309	s	aliphatic	L 644	s
aliphatic	V 318	s	aliphatic	V 654	s
aliphatic	I 332	s	aliphatic	V 668	s
aliphatic	T 334	s	aliphatic	T 670	s
aliphatic	L 389	s	aliphatic	L 799	s
aliphatic	A 399	s	aliphatic	C 809	s

[a] one-letter code of the amino acid, the number of the amino acid and origin of pseudocenter: from side-chain (s) or from peptide bond(p).

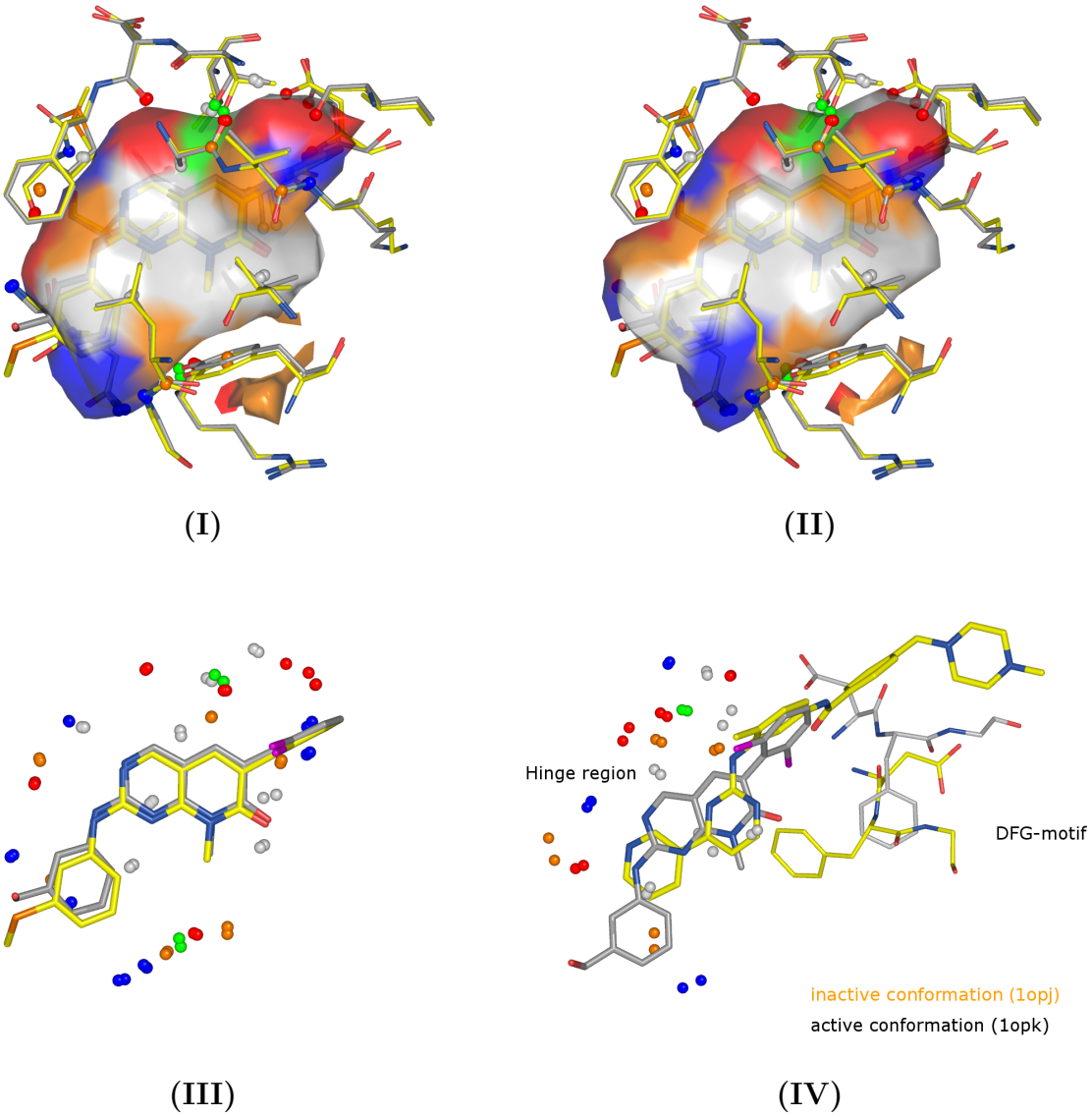


Fig. 5.15 Similarity in the ATP-binding site of c-Abl kinases of the same and different activation states. All calculations were performed with the cavity limited to the ATP pocket (residues coinciding with a 12.0Å sphere around to the hinge region). Equivalent areas in the active sites of c-Abl kinases exhibiting the same activation state are displayed in (I) to (III). The amino acids and ligands from 1opk (carbon atoms colored in yellow) and 1m52 (carbon atoms colored in white) with the matching pseudocenters and cavity surface is shown in (I) and (II). In (I) and (II) only the matching surface area from one cavity are shown (1opk and 1m52 respectively). In (III) the matching pseudocenters and bound ligands are displayed, both cavities share great portion of similarity in the active site. Inactive c-Abl (PDB code 1opj carbon atoms in yellow) and active c-Abl (PDB code 1opk carbon atoms colored in grey) c-Abl kinases show less similarity in the ATP site (IV). The matching areas in both cavities (indicated by the matched pseudocenters) are found near the adenine binding pocket and the hydrophobic region II. The conformation and spatial orientation of the DFG motif differs drastically between both states of the c-Abl kinases (DFG-in in the active conformation, DFG-out in the inactive one). Only the aspartate of the DFG motif was part of the query cavity. The flipped arrangement of the aspartate and phenylanlanine in the DFG motif can easily be visualized using the superposition based on the similarities in the ATP pocket. This superposition indicates, that Gleevec cannot bind to the active conformation (with DFG-in conformation).

5.4.8 Rationalizing the cross-reactivity of Gleevec against other kinases

In vitro screens against a panel of protein kinases have shown that STI571 inhibits besides c-Abl at least two other kinases: the stem cell receptor c-Kit (a protein tyrosine kinase) and the platelet-derived growth factor (PDGF) receptor [Capdeville et al., 2002]. Both proteins belong to the type III transmembrane receptor protein tyrosine kinase (RPTK) family. Three c-Kit structures are present in the dataset: one uncomplexed structure (PDB code 1t45), and two structures in the activated state with bound ADP (PDB code 1pkg) or STI-571 (PDB code 1t46). A comparison of the c-Abl cavities against those of all other kinases in the 263 dataset shows that the c-Kit cavities exhibit highest similarity to the c-Abl cavities. The overlapping and similar areas in both binding sites are displayed in Figure 5.16. Similarities are exhibited by the adenine region and the phosphate binding pocket. The observed similarity between the ATP-binding sites of c-Abl and c-Kit provides an explanation in structural terms for the observed cross-reactivity and the similarity of the inhibition profiles of STI-571 towards both kinases.

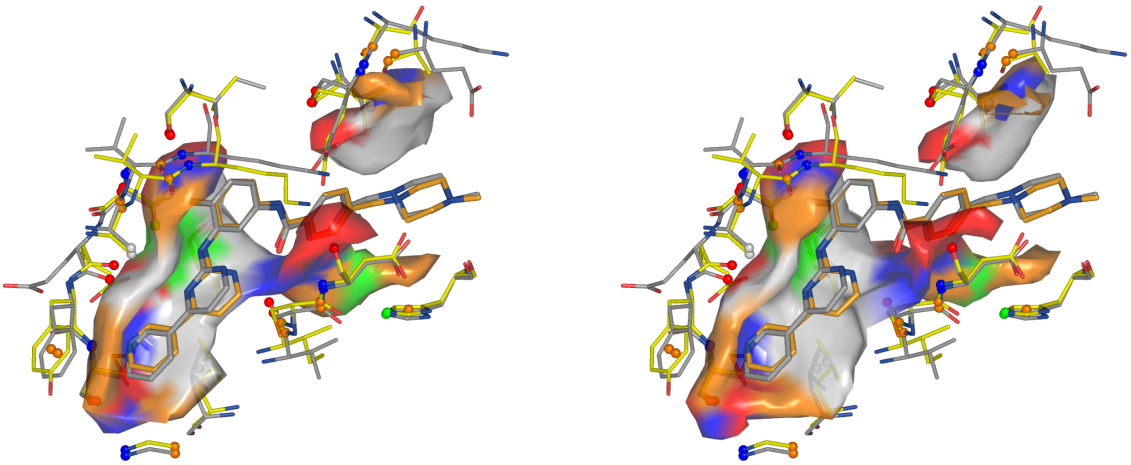


Fig. 5.16 Rationalizing the cross-reactivity of Gleevec against c-Abl and c-Kit. Similar regions in the active sites of c-Abl (PDB code 1opj) and c-Kit (PDB code 1t46) are displayed. The matching amino acids, ligands, pseudocenters are shown. Additionally, for every cavity the matching cavity surface is displayed in (I) and (II), respectively. Both kinases show a high degree of similarity in their binding sites. The pseudocenters (RMSD=0.61) and inhibitor molecules superimpose perfectly.

5.4.9 Cross relationships between unrelated kinases

Most interesting from a medicinal chemistry point of view is the structural similarity between kinases that are not closely related in sequence space. Knowledge about similar structural and physicochemical properties in their active sites can help to select kinases for cross-reactivity testing of lead compounds that are expected to perform as highly selective drugs. In order to identify unexpected similarities and dissimilarities in ATP-binding sites of the 263 kinases in our dataset, the sequence-based similarity is plotted against the computed Cavbase similarity score (R_1,R_2,R_3). Figure 5.17 plots the Cavbase similarity score R_1 faced against a normalized sequence identity score (to values between zero and one) calculated with FASTA. This plot indicates some interesting cases where a low sequence similarity is accompanied with a high Cavbase similarity, or vice versa.

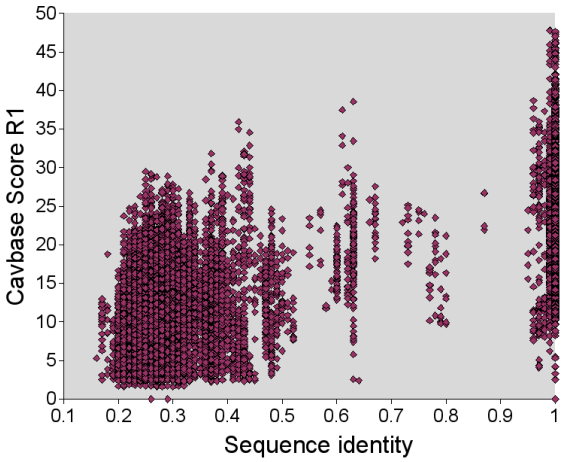
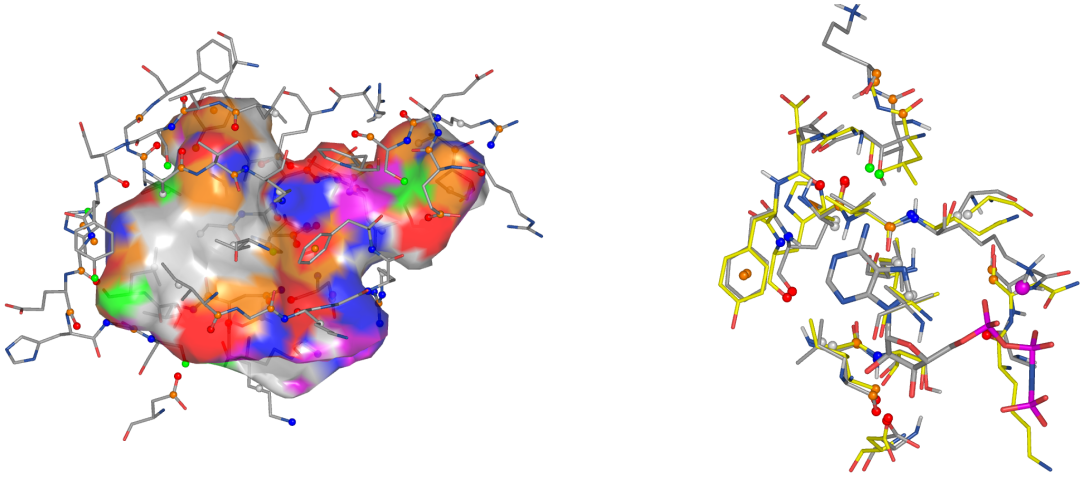


Fig. 5.17 Cavbase similarity score versus FASTA sequence identity for all kinase pairs in the dataset. The sequence similarity and Cavbase similarity index is plotted for every pairwise comparison. The calculations were performed for cavities comprising the adenine and sugar pocket together with the phosphate binding region (16.0 Å) using scoring scheme R_1. Similarity values for the self comparisons of cavities have been excluded from the dataset for reasons of clarity.

5.4.9.1 Low sequence similarity and high Cavbase similarity

Several kinase pairs exhibit a sequence similarity below 35% (0.35) and show a high Cavbase similarity $R_1 > 25.0$ (Table 5.8). For example, the tyrosine kinase of the type I TGF β receptor has pronounced similarities with kinases from the src-family, such as Hck, Csk, Lck. In Figure 5.18-II the superposition of a TGF β kinase (PDB code 1ias) and a Haemopoetic cell kinase Hck (PDB code 1ad5) is shown. The spatial similarity of both cavities encompass the adenine binding pocket, both hydrophobic sites and extends towards the phosphate binding groove. Both catalytic protein kinase domains show only a sequence identity of 26%, the Cavbase similarities matched in both binding sites are listed in Table 5.7.



(I) **(II)**

Fig. 5.18 Unexpected high similarities of type I TGF β and c-src based binding sites In (I) the complete ATP-binding site of type I TGF β cavity is shown for reference. (II) displays the large regions of type I TGF (PDB code 1ias; carbon atoms colored in grey) and Hck (PDB code 1ad5; carbon atoms colored in yellow), which are recognized as similar by Cavbase. Compared to the complete size of the TGF β cavity (I), the similar found areas are very large (indicated by the matching amino acids and pseudocenters (II)).

Tab. 5.7 As equivalent detected areas in the binding site of Hck and Type 1 TGF β.

Hck (1ad5)			**Type 1 TGF β** (1ias)		
pseudocenter type	equivalent amino acid [a]		pseudocenter type	equivalent amino acid[a]	
pi	L 273	p	pi	I 211	p
acceptor	L 273	p	acceptor	I 211	p
pi	A 293	p	pi	A 230	p
acceptor	A 293	p	acceptor	A 230	p
pi	V 294	p	pi	V 231	p
donor	K 295	p	donor	K 232	p
pi	V 323	p	pi	L 260	p
pi	K 324	p	pi	G 261	p
don_acc	T 338	s	don_acc	S 280	s
acceptor	E 339	p	acceptor	D 281	p
aromatic	F 340	s	aromatic	Y 282	s
donor	M 341	p	donor	H 283	p
acceptor	M 341	p	acceptor	H 283	p
pi	G 344	p	pi	G 286	p
donor	S 345	p	donor	S 287	p
acceptor	D 348	s	acceptor	D 290	s
acceptor	A 390	p	acceptor	K 337	p
pi	N 391	p	pi	N 338	p
aliphatic	L 273	s	aliphatic	I 211	s
aliphatic	V 281	s	aliphatic	V 219	s
aliphatic	A 293	s	aliphatic	A 230	s
aliphatic	L 393	s	aliphatic	L 340	s
aliphatic	A 403	s	aliphatic	A 350	s

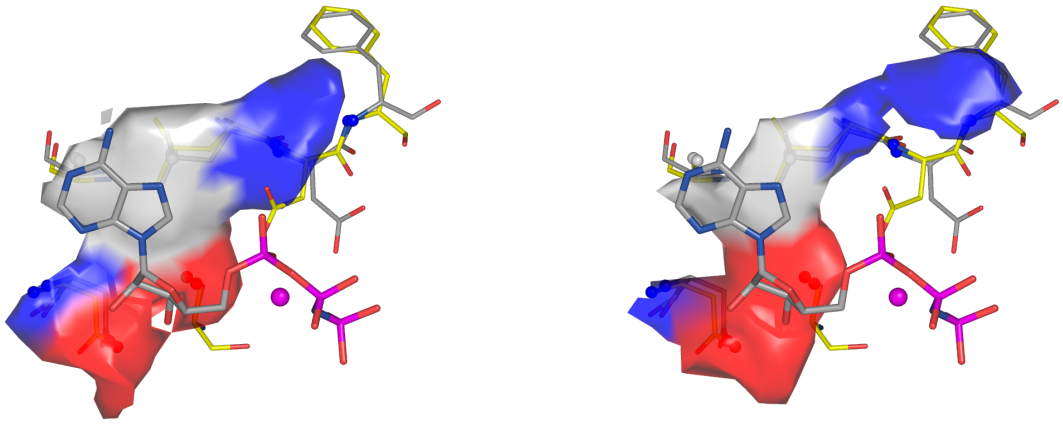
[a] one-letter code of the amino acid, the number of the amino acid and origin of pseudocenter: from side-chain (s) or from peptide bond(p).

Tab. 5.8 Unexpected similarities and dissimilarities in the binding sites of kinases. The first part of the table contains cavity pairs from kinase families which exhibit high sequence identity (> 60%) and low Cavbase similarity ($R_1 < 10.0$). Whereas the second part annotates cavity pairs which show a high Cavbase similarity ($R_1 > 25.0$) but only low sequence identity (< 35%).

SCOP superfamily	**SCOP superfamily**
low sequence identity and high Cavbase similarity	
cAMP-dependent PK, catalytic subunit	Mycobacterial protein kinase PknB, catalytic domain
cAMP-dependent PK, catalytic subunit	Death-associated protein kinase, Dap
cAMP-dependent PK, catalytic subunit	γ-subunit of glycogen phosphorylase kinase (Phk)
Death-associated protein kinase, Dap	Mutants of c-AMP that mimic Protein kinase b
Death-associated protein kinase, Dap	Pkb kinase
Lymphocyte kinase (lck)	Type I TGF β receptor R4
Haemopoetic cell kinase Hck	Type I TGF β receptor R4
Carboxyl-terminal src kinase (csk)	Type I TGF beta receptor R4
Cyclin-dependent PK, CDK2	Lymphocyte kinase (lck)
Cyclin-dependent PK, CDK2	γ-subunit of glycogen phosphorylase kinase (Phk)
Abelsone tyrosine kinase (abl)	Cyclin-dependent PK, CDK2
Abelsone tyrosine kinase (abl)	Lymphocyte kinase (lck)
γ-subunit of glycogen phosphorylase kinase (Phk)	Lymphocyte kinase (lck)
γ-subunit of glycogen phosphorylase kinase (Phk)	Sky1p
high sequence identity and low Cavbase similarity	
Cyclin-dependent PK, CDK2	Cyclin-dependent PK, CDK2 + Cyclin
MAP kinase p38α	MAP kinase p38γ
3-phosphoinositide dependent protein kinase-1 Pdk1	Choline kinase
cAMP-dependent PK, catalytic subunit	Mutants of c-AMP that mimic Protein kinase b
Insulin receptor	Insulin-like growth factor 1 receptor

5.4.9.2 High sequence similarity and low Cavbase similarity

Besides the above-discussed cases arising from different activation states of the same kinase that shows identity in sequence space however low similarity in cavity space, there are several subfamilies that are closely related in sequence space but strongly deviate in cavity space (Table 5.8). As an example, the MAP kinases p38α and p38γ display a high sequence similarity (both kinases are found in one sequence space cluster; sequence identity is 63%), whereas only modest Cavbase similarity is exhibited. In view of this disparity, it is interesting to note that Vieth and coworkers [Vieth et al., 2004] could show, that these two related families differ with respect to the inhibition profiles of small molecule inhibitors addressing these kinases. Binding site similarities exhibited by cavities of two members of these families are shown in Figure 5.19.



(I) **(II)**

Fig. 5.19 Similar areas in the binding sites of MAP kinase p38γ (PDB code 1cm8) and MAP kinase p38α (PDB code 1p38). The matching area of both cavities is rather small and located next to the hinge region. Displayed are the amino acids, pseudocenters, bound ATP and one cavity surface for 1cm8 (I) and 1p38 (II), respectively.

5.5 Conclusions and outlook

In this study, a shift from the comparison of individual cavities against the entire cavity database towards the clustering of large datasets of cavities has been performed. This step was facilitated by a significant speed-up of the cavity similarity calculations

and the application of clustering tools. The results of the cluster analysis revealed that similarities and differences in the binding sites of proteins could be utilized to obtain relevant classification results, even across highly homologous protein families. The obtained classification complements classification schemes based on similarities in sequences, fold patterns, or inhibition profiles of small molecules.

An advantage of this new approach is the concentration on the functionally relevant areas of the proteins. Additionally, the analysis can be performed in a completely automated fashion and is not limited by the number of input structures used.

The calculation of a relevant 3D-alignment for multiple protein structures is a non-trivial task, especially if the structures do not show high degree of sequence similarity or do exhibit partial structural flexibility. The superposition and alignment of kinases based on similarities exhibited in their active sites allows for an efficient way to cope with this situation. The obtained superpositions can also be used to further characterize binding sites, e.g to reveal favorable interaction pattern for certain ligand atoms (hotspot analysis) in the binding pockets, or to provide a reasonable starting point for ligand alignment during a 3D-QSAR analysis.

The classification of the carbonic anhydrases as a rather rigid and uniform protein family revealed good classification results. Cavbase provides a classification on the subfamily level and even detects differences in one CA subfamily that arise from one residue intimately involved in catalysis. Even in a versatile and diverse protein family such as the protein kinases, the classification results are in agreement with other classification schemes. If two proteins have a high sequence similarity they generally also show a high Cavbase similarity. Additionally to these intuitively expected results, the Cavbase classification reveals interesting relationships across kinases, such as CDKs of different activation states, on the similarities between protein kinases, that are not related in sequence space.

6 Subtaschen gesteuertes Optimieren und *de novo-Design* von Inhibitoren durch Ähnlichkeitsanalysen in Proteinbindestellen

6.1 Die SARS Coronavirus M^{pro} als Target für die Suche nach antiviralen Arzneistoffen

Im Frühjahr 2003 kam es von Südchina ausgehend zu einer Epidemie-artigen Verbreitung einer atypischen Lungenentzündung: dem *Schweren Akuten Respiratorischen Syndrom (Severe Acute Respiratory Syndrome)* (SARS), [Anand et al., 2003]. Der Krankheitsverlauf von SARS ist charakterisiert durch hohes Fieber, schweren Husten und Atemnot. Die Letalität liegt je nach betroffener Region zwischen 6 und 20% (WHO, Stand 07/2003). Die Erkrankung wird durch eine neues beim Menschen bisher noch nicht beobachtetes Coronavirus hervorgerufen [Fouchier et al., 2003; Kuiken et al., 2003]. Die Coronaviridae, die Familie der Coronaviren, zählen zu den RNA-Viren. Ihr Erbgut besteht aus einem einzigen, bis zu 30 000 Nukleotiden langen, einsträngigen RNA-Molekül. Vor der Entdeckung des SARS-Erregers waren bereits zwei andere humanpathogene Coronaviren bekannt. Diese werden für etwa 15 bis 30 Prozent aller Erkältungskrankheiten verantwortlich gemacht; sie sind jedoch vergleichsweise harmlos. Coronaviren besitzen ein großes Replikase Gen, das zwei überlappende Polyproteine kodiert, die sowohl der Virusreplikation als auch in der Virustranskription involviert sind. Funktionelle Polypeptide werden durch proteolytische Spaltung der Polyproteine hauptsächlich durch eine Protease M^{pro} (main protease M^{pro}, auch 3C-ähnliche Protease, $3CL^{pro}$) freigesetzt [Ziebuhr et al., 2000]. Diese Protease stellt wegen ihrer Wichtigkeit im viralen Lebenszyklus ein interessantes Target zur Entwicklung von antiviralen Arzneistoffen dar. Wie schon in Kapitel 4.1.1 erwähnt, besitzt die CoV M^{pro} eine Serinprotease ähnliche Faltung (Domäne I und Domäne II) mit einer zusätzlichen helikalen Domäne (Domäne III). Die SARS CoV M^{pro} weist eine katalytische Diade aus einem Cystein und einem Histidin auf, wobei die Bindestelle an der Spalte zwischen Domäne I und Domäne II lokalisiert ist (siehe Abbildung 4.1). Das Faltungsmuster der

M^{pro} Proteasen verschiedener verwandter Viren ist sehr ähnlich, wie zum Beispiel bei dem humanen Coronavirus (HCoV), dem Transmissiblen Gastroenteritis Virus (TGEV) und dem Maus Hepatitis Virus [Yang et al., 2003].

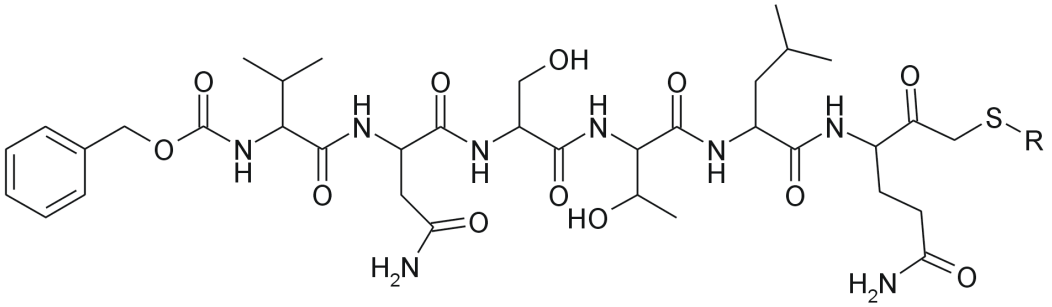
6.2 Ähnlichkeitssuche mit Cavbase

In diesem Abschnitt soll untersucht werden, ob man Ähnlichkeiten in der Bindetasche der SARS CoV M^{pro} zu anderen Proteinen nutzen kann, um neue Ideen für das Design antiviraler Proteasehemmer zu gewinnen. Um dieser Frage nachzugehen wird die Bindetasche der TGEV M^{pro} und der SARS CoV M^{pro} einer Cavbase Ähnlichkeitsanalyse unterzogen. In Kapitel 4.1.1 wurden bereits die Ergebnisse der Ähnlichkeitsanalyse mit der SARS CoV Bindetasche zum Auffinden von funktional ähnlichen Proteinen vorgestellt. Hier soll der Schwerpunkt auf der Analyse der SARS CoV M^{pro} Subtaschen liegen. Proteasen spalten Polypeptidketten. Um ein Substrat spezifisch erkennen zu können, besitzen Proteasen auf ihrer Oberfläche mehrere Bindungstaschen, die strukturell komplementär zu den Seitenketten des Substrats sind. Eine erprobte Strategie im Design von Proteaseninhibitoren ist die Abwandlung von Seitenketten des Substratpeptides, um eine stärkere oder spezifischere Bindung an das Protein zu erzielen. Als Suchanfrage werden die Subtaschen der SARS CoV M^{pro} bzw der TGEV M^{pro} als Referenz verwendet. Durch die Berücksichtigung von Subtaschen ist man in der Lage, Informationen über lokale Ähnlichkeiten mit Bindetaschenbereichen funktional nicht verwandter Proteine zu erhalten. Aus Ligandfragmenten, die in diesen so entdeckten Taschen eine ähnliche Proteinumgebung vorfinden, können neue Ideen für bioisostere Baugruppen gewonnen werden, die sich dann für das Design neuer Inhibitoren nutzen lassen. Für die Subtaschenanalyse werden folgende kristallographisch aufgeklärte M^{pro} Proteinstrukturen verwendet:

- TGEV M^{pro} [PDB Code: 1p9s, 1p9u]
- SARS CoVM^{pro} [PDB Code: 1uj1, 1uk3, 1uk2, 1uk4]

Im Folgenden werden die Ergebnisse der Cavbase Ähnlichkeitsanalysen anhand der Protease Strukturen, die einen Substrat-analogen peptidischen Inhibitor (**19**) gebunden haben (PDB Code 1p9u und 1uk4), vorgestellt [Anand et al., 2003; Yang et al., 2003] (siehe Abbildung 6.1). Es wird jeweils die komplette Bindetasche wie auch die

verschiedenen Subtaschen als Anfragetaschen in der Ähnlichkeitsanalyse verwendet. Die TGEV M^{pro} Struktur besteht aus drei Dimeren, jedes Dimer ist aus zwei Protome-



19

Abb. 6.1 Hexapeptidischer Chloromethylketon (CMK) Inhibitor von CoV M^{pro} (19). Der Inhibitor mit der Sequenz Cbz-Val-Asn-Ser-Thr-Leu-Gln-CMK ist im Komplex mit der TGEV CoV M^{pro} (PDB Code 1p9u) und der SARS CoV M^{pro} (PDB Code 1uk4) gelöst. Die N-terminale Benzyloxycarbonyl-Schutzgruppe (Cbz) ist in den gelösten Kristallstrukturen nicht sichtbar, für sie ist keine Dichte definiert.

ren zusammengesetzt. Für die Ähnlichkeitsanalyse wird die Bindetestelle benutzt, die durch die Proteindomänen E und F aufgebaut ist [Anand et al., 2003].

Die SARS CoV Struktur liegt im Kristall und in Lösungen ab einer gewissen Konzentration (>1mg/ml) als symmetrisches Dimer (Protomer A und B) vor. Für beide Protomere wird jeweils eine Substratbindestelle detektiert. Yang et al. haben Kristallstrukturen der SARS CoV M^{pro} bei den pH-Werten 6.0, 7.6 und 8.0 vermessen [Yang et al., 2003]. Die Protomere der Kristallstrukturen, die bei den beiden höheren pH-Werten gelöst wurden, sind zueinander sehr ähnlich. Auch die Bindestelle des Protomers A des bei pH 6.0 kristallisierten Enzyms liegt in einer ähnlichen Konformation wie die Substratbindetaschen der anderen Protomere vor. Die Bindetasche des Protomers B zeigt eine andere, inaktive Konformation. Durch konformative Veränderungen wird der Substratbindebereich blockiert und das sogenannten O^--Loch (*oxyaninon hole*) wird nicht ausgebildet. Bei einem pH-Wert von 6.0 ist also nur eine der beiden Substratbindestellen in der Lage, das Substrat in einer für die Reaktion produktiven Form zu binden. Das spiegelt sich auch in den kinetischen Daten wider, bei pH 6.0 zeigt die SARS CoV M^{pro} nur 50% ihrer Aktivität. Die SARS CoV Kristallstruktur mit gebundenem Substratanalogon (**19**) wurde bei pH 6.0 kristallisiert. Sie zeigt ebenfalls ein aktives und ein inaktives Protomer. Beide Protomere haben den Inhibitor über eine kovalente Verknüpfung zu Sγ von Cys 145 über ihre CMK-Methylengruppe

gebunden. Eine überraschende Konformation des Inhibitors in der Bindetasche ist zu beobachten. Abbildung 6.2 zeigt die zu erwartende Besetzung der Subtaschen in der TGEV M^{pro} (I) und die nicht zu erwartende in der SARS M^{pro}. Darüber hinaus werden die Subtaschen in den beiden Protomeren der SARS M^{pro} verschieden besetzt. Hier ist der Bindungsmodus für das Protomer A (aktive Konformation) aufgelistet (siehe Abbildung 6.2-II).

- Leucin P2 bindet nicht in die S2 Tasche (in die Nähe von Asp 187), sondern verbleibt Solvens-exponiert
- Threonin P3 bindet in die S2 Tasche. Die Spezifität der SARS CoV M^{pro} für S2 Reste ist nicht so hoch wie bei anderen CoV M^{pro} (beschränkt auf Leucin)
- Asparagin P5 bindet in die S4 Subtasche

Der Bindungsmodus für das inaktive Protomer B weist ebenfalls einen unerwarteten Bindungsmodus auf:

- Der Inhibitor bindet nicht in die S1 Tasche
- Glutamin P1 ist Solvens-exponiert
- Leucin P2 und Serin P4 zeigen in ihre entsprechenden Subtaschen

Der peptische Ligand nimmt in der SARS CoV Struktur wie oben beschrieben eine für Cysteinproteasen untypische Besetzung der Spezifitätstaschen ein (siehe Abbildung 6.2-II). Sie werden nicht von zu erwartenden Seitenketten des Substratanalogen Inhibitors adressiert. In der TGEV M^{pro} Kristallstruktur wird ein Bindungsmodus des peptidischen Inhibitors in der zu erwarteten Besetzung der Spezifitätstaschen gefunden (siehe Abbildung 6.3-I). Wie unterschiedlich sind sich die TGEV M^{pro} und die SARS CoV M^{pro} in ihren Bindetaschen? Die Überlagerung mit Cavbase zeigt, daß große Bereiche in der Bindetasche als ähnlich angesehen werden können, aber auch Bereiche in beiden Bindetaschen vorliegen, die unterschiedlich sind. So nimmt die Seitenkette des katalytischen Cysteins in beiden Strukturen eine unterschiedliche Position ein. Die Konformationen der beiden Inhibitoren sind verschieden, dadurch nehmen die Cysteine in beiden Strukturen eine unterschiedliche Position in der Bindetasche nach der Überlagerung beider Taschen ein (siehe Abbildung 6.3). Strukturelle Unterschiede in beiden Bindetaschen finden sich vor allem im Bereich der Subtaschen S2, S3, S4 und S5 (siehe Abbildung 6.4,

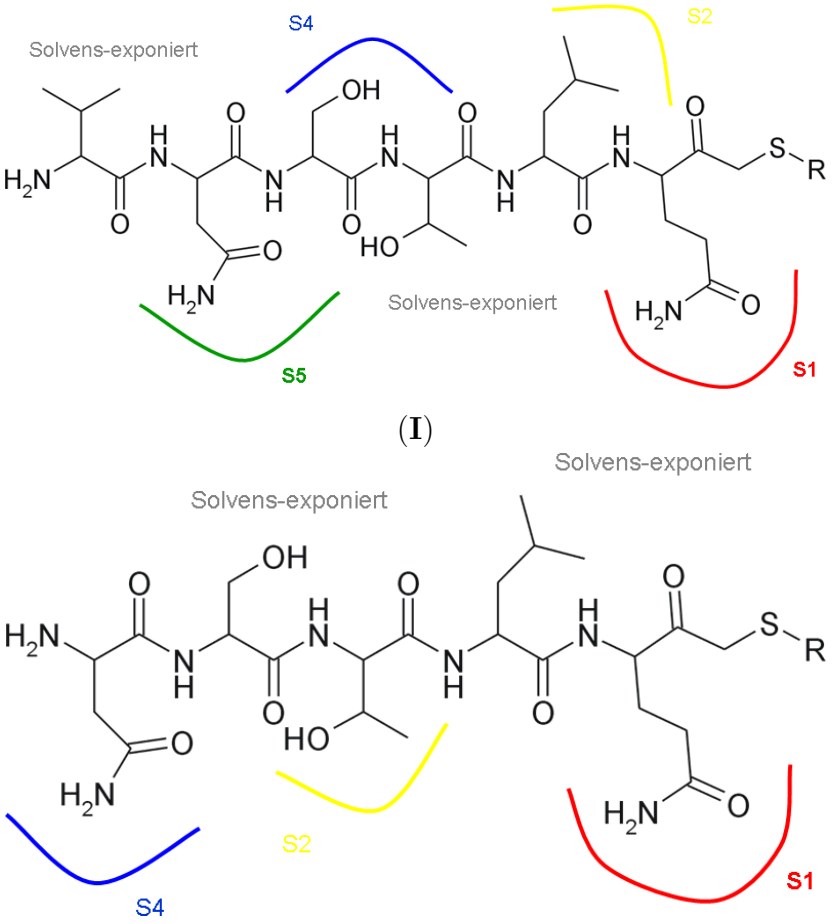


Abb. 6.2 Schematische Darstellung des Bindungsmodus des peptidischen CMK-Inhibitors in TGEV M^{pro} (PDB Code 1p9u) und SARS CoV M^{pro} (PDB Code 1uk4). In (I) ist der erwartete Bindungsmodus des CMK-Inhibitors in TGEV dargestellt, während (II) die unerwartete Besetzung in der SARS CoV M^{pro} (Protomer A) zeigt. In der SARS CoV M^{pro} Struktur ist zusätzlich zur Cbz-Schutzgruppe keine Dichte für das N-terminale Valin definiert.

I). Die Überlagerung der Inhibitoren zeigt ebenfalls den unterschiedlichen Bindungsmodus, während beide P1-Glutaminreste räumlich noch ungefähr überlagert werden, läßt sich die unterschiedliche Besetzung der anderen Subtaschen gut erkennen. Zur Ähnlichkeitssuche mit Cavbase werden die einzelnen Subtaschen der TGEV M^{pro} und der SARS M^{pro} verwendet (siehe Tabelle 6.1) Sie wurden visuell beurteilt und manuell ausgeschnitten, anschließend als eigenständige Bindetaschen in Cavbase abgelegt.

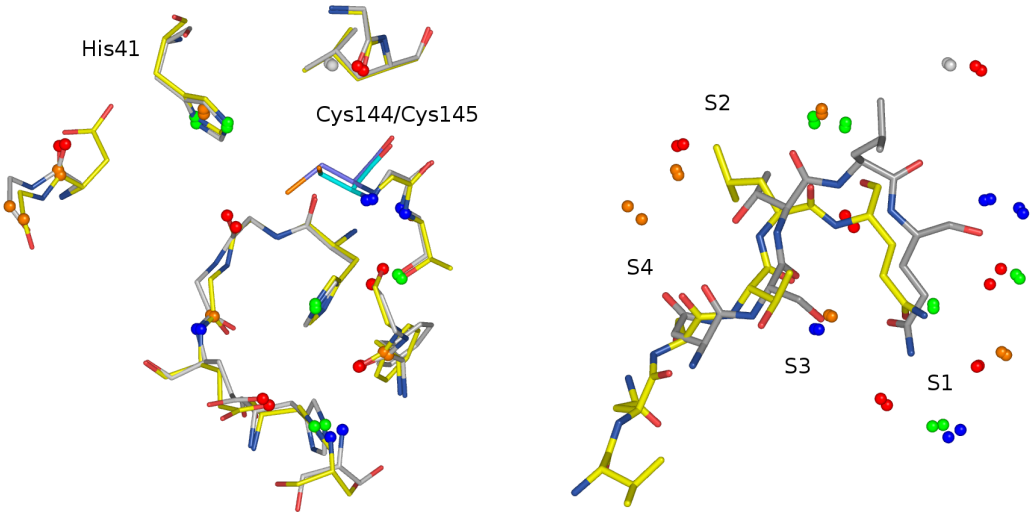


Abb. 6.3 Vom Vergleichsalgorithmus als ähnlich erkannte Bereiche in der Bindetasche von TGEV M^{pro} (PDB Code 1p9u) und SARS CoV M^{pro} (PDB Code 1uk4). In (I) sind die als ähnlich erkannten Aminosäuren und Pseudozentren der TGEV M^{pro} (Kohlenstoffe in gelb gefärbt) und der SARS CoV (Kohlenstoffe in weiß gefärbt) dargestellt. Das katalytische Cystein ist nicht Teil des gefundenen Musters. Der Inhibitor bildet mit den Thiolgruppen des Cysteins 145 eine kovalente Bindung aus (Bindung nicht dargestellt). Bedingt durch die unterschiedlichen Konformationen der Liganden sind die Thiolgruppen räumlich voneinander entfernt und werden nicht als ähnlich erkannt (1p9u Cys144, Kohlenstoffe in cyan; 1uk4 Cys145 Kohlenstoffe in blau). Aus Gründen der Übersichtlichkeit werden einige Aminosäuren nur durch ihre Hauptkettenatome und die Bindung des Cysteins zur Chlormethylketo-Gruppe des Inhibitors nicht dargestellt. In (II) sind die gebundenen Liganden (Inhibitor in 1p9u, Kohlenstoffe in weiß gefärbt; Inhibitor in 1uk4 Kohlenstoffe in gelb gefärbt) und die als ähnlich gefundenen Pseudozentren gezeigt. Deutlich ist die unterschiedliche Besetzung der Subtaschen und die 'Verschiebung' der Seitenketten des Inhibitors zu erkennen.

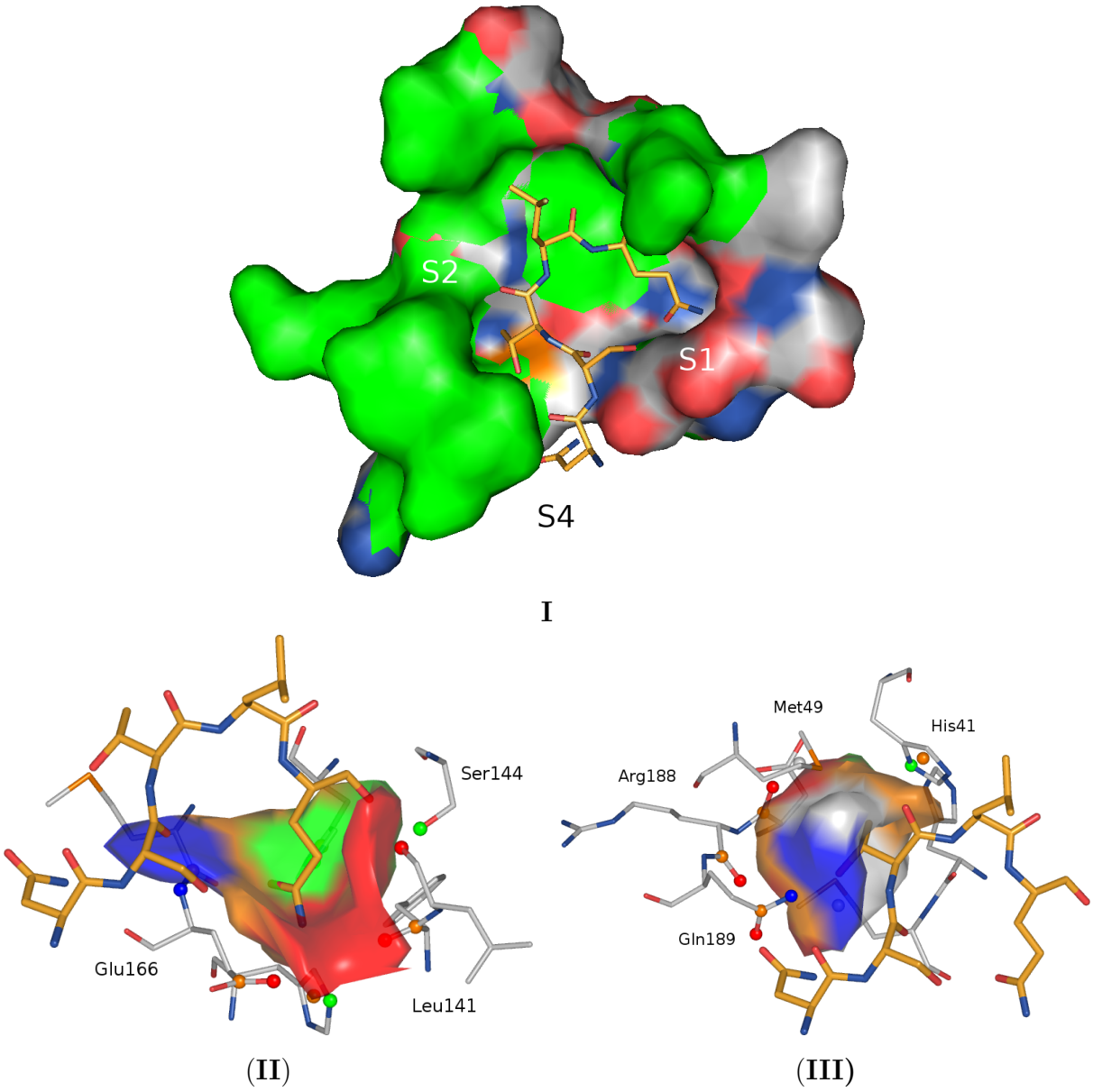


Abb. 6.4 Unterschiedliche räumliche Gestalt der TGEV M^{pro} (PDB Code 1p9u) und SARS CoV M^{pro} (PDB Code 1uk4) und Ausmaß der S1- und S2-Subtasche der SARS CoV M^{pro}. In (I) sind Oberflächenbereiche, die in beiden Strukturen eine ähnlichen Raumbereich einnehmen (nach Atomtypen gefärbt) und räumlich unterschiedliche Oberflächenbereiche (in grün gefärbt) der TGEV M^{pro} und der SARS M^{pro} Bindetaschen gezeigt. Der Chlormethylketon-Inhibitor aus der SARS CoV M^{pro} ist zur Orientierung mit dargestellt. Strukturelle Unterschiede zwischen beiden Strukturen treten vor allem im Bereich der S2 und S4 Tasche auf. In (II) und (III) sind die S1-Tasche und S2-Tasche der SARS CoV M^{pro} dargestellt. Die Auswahl der Pseudozentren wurde manuell vorgenommen, um eine vollständige Beschreibung der Subtaschen zu erreichen. Die S1-Tasche hat einen eher polaren Charakter, während die S2-Tasche einen aliphatisch-aromatischen Charakter aufweist.

Tab. 6.1 Aminosäurezusammensetzung der einzelnen Subtaschen von TGEV M^{pro} (PDB Code 1p9u) und SARS CoV M^{pro} (PDB Code 1uk4)

Subtasche	CavID	Zusammensetzung der TGEV M^{pro} (PDB Code 1p9u) Subtaschen nach [Anand et al., 2003]
S1	—	Leu164, Glu165, His171
S2	—	Leu164, Ile51, Thr47, His41, Tyr53
S3	—	zum Solvens exponiert
S4	—	Leu164, Leu166, Gln191
S5	—	Gly167, Ser189, Gln191
S6	—	zum Solvens exponiert
Subtasche	CavID	Zusammensetzung der TGEV M^{pro} (PDB Code 1p9u) Subtaschen nach visueller Inspektion
S1	1p9u.89	Phe139 His162 Leu164 Glu165 His171
S2	1p9u.88	His41 Thr47 Leu164 Asp186 Gln187 Ile51
S3	—	zum Solvens exponiert
S4	1p9u.87	Leu164 Leu166 Gln187 Ser189 Gln191
S5	1p9u.86	Ser189 Met190 Gln191 Gly167 Leu166
Subtasche	CavID	Zusammensetzung der SARS CoV M^{pro} (PDB Code 1uk4) Subtaschen nach visueller Inspektion
S1	1uk4.99	Phe140 Leu141 Ser144 His163 Met165 Glu166 His172
S2	1uk4.98	His41 Met49 Met165 Asp187 Arg188 Gln189
S3	—	Solvens exponiert

Tab. 6.2 Gruppen von Ligandfragmenten, die in ein ähnliches Umfeld wie in der die S1-Tasche der SARS-CoV M^{pro} binden.

Gebundenes Ligandfragment	Anzahl
Liganden mit Aminogruppen, basische Verbindungen	16
Aliphatische Liganden	4
Liganden mit Phosphatgruppen	19
Liganden mit Zuckern	21
Aromatische Liganden	6
Hämoglobin-artige Liganden	8
Liganden mit Alkohol- oder Säurefunktion	13

6.3 Ähnlichkeitssuche mit der SARS CoV M^{pro} Subtaschen

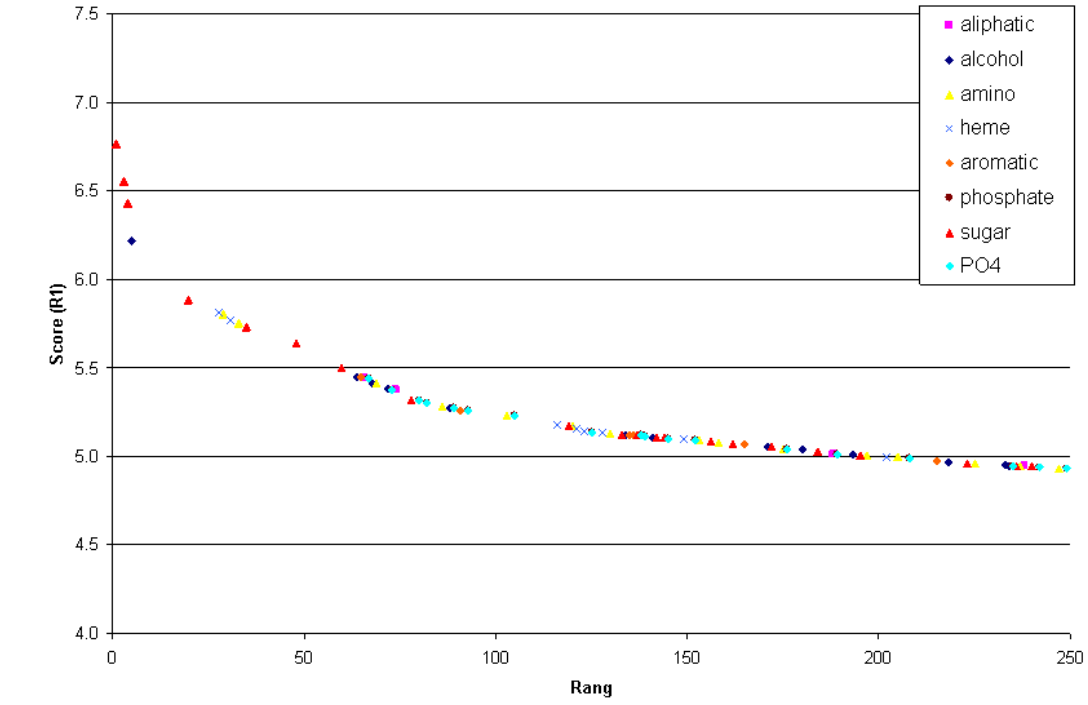
In dieser Arbeit sollen neue Ideen für das Design von Wirkstoffen, die potentiell gegen SARS CoV M^{pro} binden, vorgeschlagen werden. Deshalb werden in diesem Abschnitt nur die Ergebnisse der Suche mit den SARS CoV Subtaschen vorgestellt, da die SARS CoV M^{pro} das eigentliche Zielprotein darstellt. Als Suchtasche wurden die Subtaschen, wie sie in Tabelle 6.1 definiert werden, verwendet. Die **S1**-Tasche der SARS CoV M^{pro} (siehe Abbildung 6.4, II) wurde gegen 9352 Bindetaschen, die einen Arzneistoff-ähnliches Molekül gebunden haben, verglichen. Die S1-Subtasche umfasst 12 Pseudozentren und besitzt polaren Charakter. Cavbase liefert als Ergebnis eine Liste von Bindetaschen geordnet nach der Bewertungsfunktion R_1. Die ersten 250 Ränge wurden visuell inspiziert. Liganden wurden danach beurteilt, ob sie ähnliche Bereiche in der Bindetasche binden und die Subtasche gut ausfüllen, d. h. mit der jeweiligen Aminosäure des Inhibitors räumlich überlagern. Die gefundenen Ligandfragmente, die in die Subtasche binden, wurden in sieben Gruppen nach ihrer chemischen Charakter klassifiziert: Liganden mit Alkohol-/Säuregruppe, mit aliphatischem Rest, mit aromatischem Rest, mit Phosphatgruppen, mit basischen Gruppen oder Hämoglobin-artige Liganden. Damit gelingt eine Charakterisierung der Bindetasche aufgrund von Ligandfragmenten, die in ähnliche Bindetaschenbereiche in anderen Strukturen binden. In Abbildung 6.5 ist eine graphische Darstellung der ersten 250 Ränge für die S1- und S2-Tasche gezeigt.

Tab. 6.3 Gruppen von Ligandfragmente, die in ein ähnliches Umfeld wie die S2-Tasche der SARS-CoV M^{pro} binden

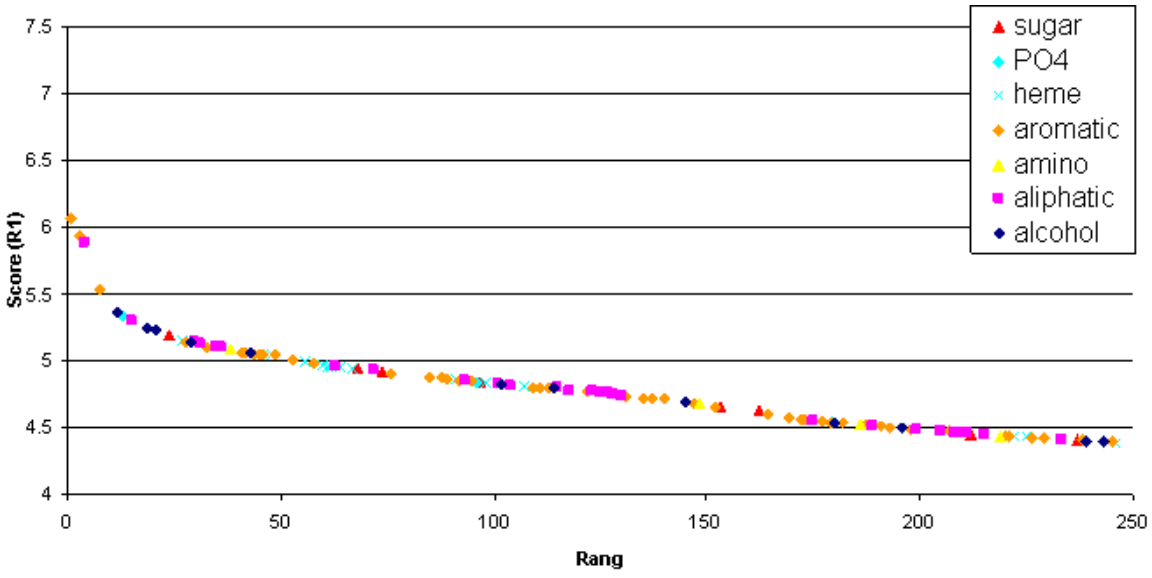
Gebundenes Ligandfragment 1uk4.97	Anzahl
Liganden mit Aminogruppen, basische Verbindungen	4
Aliphatische Liganden	28
Liganden mit Phosphatgruppen	2
Liganden mit Zuckern	8
Aromatische Liganden	49
Hämoglobin-artige Liganden	15
Liganden mit Alkohol- oder Säurefunktion	12

Im Fall der S1-Subtasche werden unter den ersten 250 Rängen 87 Bindetasche mit Ligandfragmenten gefunden, die mit der Seitenkette des Glutamins des Inhibitors überlagern. Die Mehrzahl der gefundenen Liganden hat einen polaren Charakter, so werden am häufigsten zuckerhaltige Bausteine, Fragmente mit Phosphatgruppen und mit basischen Gruppen (Amino-, Guanidino-gruppen) entdeckt. Substrate, die von CoV M^{pro} gespalten werden, weisen an der P1-Stelle immer ein Glutamin auf [Ziebuhr et al., 2000]. In Abbildung 6.6-I und II sind ausgewählte Liganden, die bei der Cavbase Ähnlichkeitssuche mit den Subtaschen gefunden wurden, dargestellt.

Die Analyse der S2-Tasche zeigt, daß die Ligandfragmente einen aromatisch-aliphatischen Charakter besitzen (wie z.B. Phenyl-, Imidazol-, Pyrimidin-Ringe). Ein Großteil dieser Fragmente entstammt anderen Proteaseinhibitoren. Es werden aber auch 12 Liganden mit Serin/Threonin-ähnlichen Alkoholgruppen gefunden, die gut mit dem in der SARS CoV M^{pro} gebundenem Threonin überlagern. In Abbildung 6.6-III und IV sind zwei Beispiele für gefundene Bindetasche mit aromatischen und aliphatischen Ligandfragmenten gezeigt.



(**I**)



(**II**)

Abb. 6.5 Die ersten 250 Ränge der Ähnlichkeitsanalyse mit der S1-Tasche und der S2-Tasche der SARS CoV M^{pro}. Beide Abbildungen zeigen die Cavbase-Bewertung (Scoringfunktion R_1) aufgetragen gegen die gefundene Platzierung. Bindetaschen, deren Liganden nach der Cavbase Überlagerung in die jeweilige SARS CoV M^{pro} Subtasche (S1-Tasche (I) oder der S2-Tasche (II)) zeigen, sind als farbige Symbole dargestellt. Die Art und Farbe des Symbols entspricht je nach dem chemischen Charakter des gebundenen Liganden einer der sieben verschiedenen Kategorien.

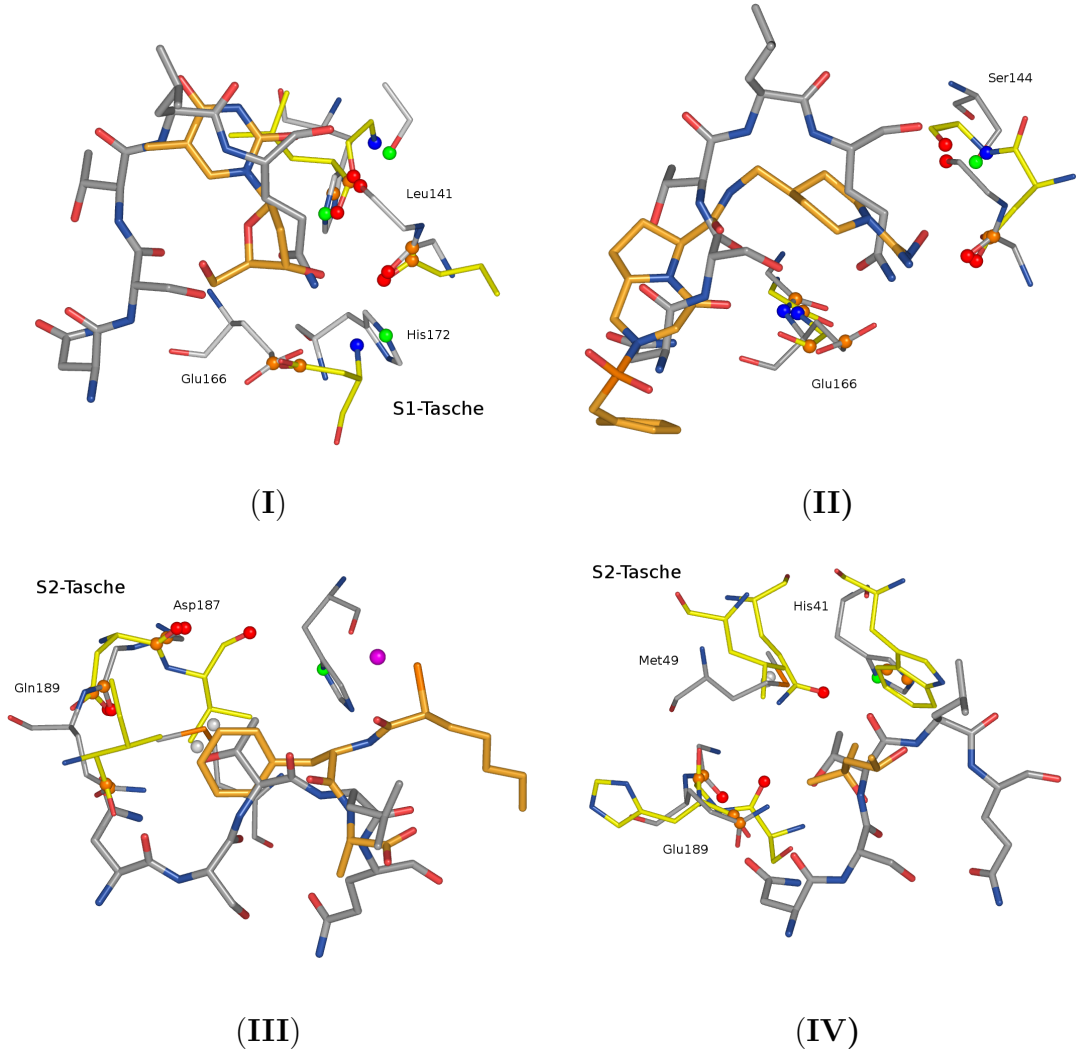


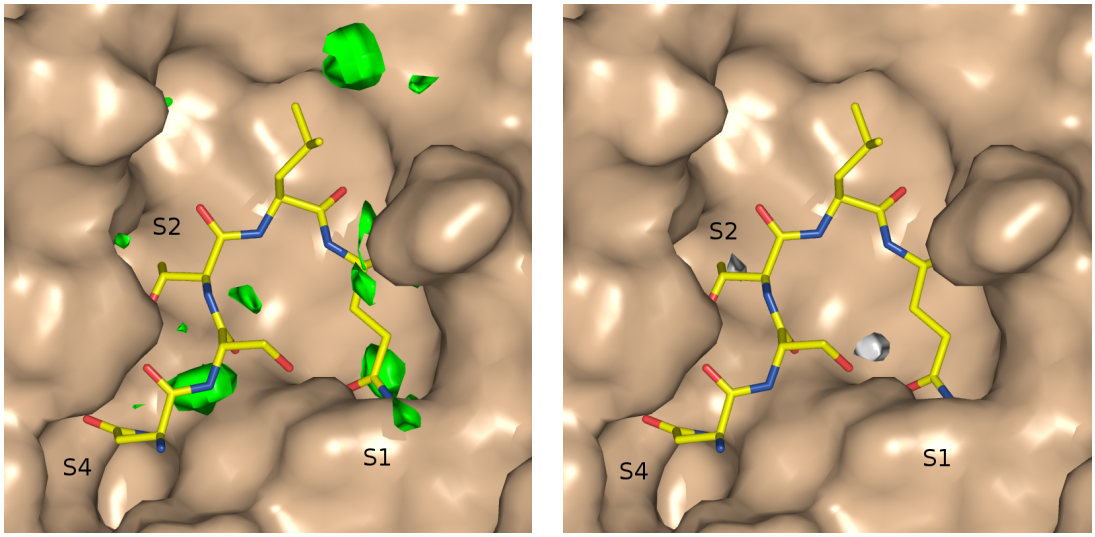
Abb. 6.6 Ligandenfragmente, die in ähnliche Subtaschen wie die SARS M^{pro}-Subtaschen binden. In (I) sind die ähnlichen Bereiche der SARS M^{pro} S1-Subtasche und der Glucose-1-Phosphat Thymidylyltransferase (PDB code 1g0r) mit gebundenem Thymidin dargestellt. Der Deoxyribosezucker des Thymidins überlagert mit dem Glutaminrest des SARS M^{pro} Inhibitors. Die Struktur wird auf Rang 2 gefunden. In (II) sind die als ähnlich gefundenen Bereiche in der S1-Subtasche und einer Thrombinstruktur (PDB Code 1jwt; Kohlenstoffe in gelb gefärbt) mit gebundenem Inhibitor gezeigt. Einige Seitenketten, die nicht durch Pseudozentren repräsentiert werden, sind aus Gründen der Übersichtlichkeit nicht dargestellt. Auf dem zweiten Rang der Ähnlichkeitssuche mit der S2-Tasche wird eine Thermolysin Bindetasche (PDB Code 1qf1; Kohlenstoffe in gelb gefärbt) gefunden. Der Phenylring des Thermolysin-Inhibitors bindet in ein ähnliches physikochemisches Umfeld wie der Threoninrest des M^{pro} Inhibitors. In (IV) ist ein Beispiel für eine Bindetasche mit einem Liganden gezeigt, der eine aliphatische Gruppe in S2 platziert. Auf dem 32. Rang wird eine Bindetasche eines Immunglobulins mit gebundenem 2-Methyl-2,4-Pentandiol gefunden (PDB Code 1fzk; Kohlenstoffe in gelb gefärbt). Die Dimethylgruppe des Liganden nimmt eine ähnliche Position wie das Threonin in der SARS-Strutur ein.

6.4 *Hotspot* Analyse der SARS CoV M^{pro}

SuperStar [Verdonk et al., 1999, 2001] ist ein Verfahren, um Interaktionsstellen in Proteinbindetaschen zu identifizieren, indem Wahrscheinlichkeitsdichten für das Auftreten bestimmter Ligandatomtypen in der Bindetasche in Form sogenannter *Hotspots* visualisiert werden (siehe Kapitel 3.4.1). Es handelt sich dabei um einen wissensbasierten Ansatz, da nur experimentelle Informationen über nichtbindende Wechselwirkungen, entnommen aus einer Vielzahl von Kristallstrukturen, verwendet werden. Für die Analyse der SARS CoV M^{pro}-Bindetasche wurden folgende Sonden aus IsoStar verwendet:

- Amino-Stickstoff (amino NH)
- Alkohol-Sauerstoff (alcohol oxygen)
- Carbonyl-Sauerstoff (carbonyl oxygen)
- positiv geladener Stickstoff (charged NH)
- ungeladener Stickstoff (uncharged NH)
- aromatischer Kohlenstoff (aromatic CH)
- aliphatischer Kohlenstoff (aliphatic CH)

Die *Hotspot*-Analyse für eine aromatische Sonde zeigt keine besonders günstigen Wechselwirkungsbereiche. Auch für die Amino-Stickstoff-Sonde (mögliche Akzeptoren auf Proteinseite) wurden keine stark favorisierten Feldbeiträge gefunden (nicht dargestellt). *Hotspots* für eine Alkohol-Sauerstoff-Sonde werden u.a. in der S1-Tasche detektiert (siehe Abbildung 6.7). Dies steht im Einklang mit den Ligandfragmenten, die bei der Cavbaseähnlichkeitsanalyse entdeckt wurden. Im Bereich der S2 Taschen werden hydrophobe Feldbeiträge entdeckt, was ebenfalls mit den Daten über gebundene Fragmente aus der Cavbaseähnlichkeitsanalyse übereinstimmt und den hydrophoben Charakter dieser Tasche gut wiedergibt. (siehe Abschnitt 6.3).



(**I**) (**II**)

Abb. 6.7 *Hotspots* in der SARS CoV M^{pro}-Bindetasche berechnet mit dem Programm Superstar. Zur Orientierung ist auch die kristallographisch bestimmte Bindungsgeometrie des Substratanalogons (gelb) gezeigt. (I) 'Hot Spots' für einen Alkohol-Hydroxylgruppe, die sowohl als Wasserstoffbrückendonor wie -akzeptor wirken kann, konturiert auf dem Wahrscheinlichkeitsniveau 8 (grün), d.h. die Wahrscheinlichkeit an dieser Stelle eine Alkoholgruppe zu finden ist achtmal höher als die durchschnittliche Wahrscheinlichkeit. *Hotspots* finden sich in der Nähe der Amidgruppe des Glutamins in der P1-Tasche. (II) Günstige Wechselwirkungsbereiche für ein aliphatisches Sondenatom konturiert auf das Wahrscheinlichkeitsniveau 10 (weiß). In der Nähe des Threonins, das in die S2-Tasche bindet, findet sich ein aliphatischer *Hotspot*. Der Solvensradius zur Berechnung der Solvens-zugänglichen Oberfläche wurde auf 1.2 Å gesetzt.

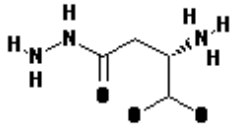
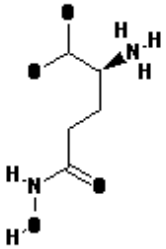
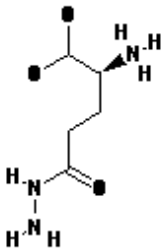
6.5 Docking Studien an den SARS CoV M^{pro} Subtaschen

Für die Dockingstudie wurde das Programm FlexX (Version 1.13) [Rarey et al., 1996] benutzt. Die Koordinaten des gebundenen Inhibitorpeptids der SARS CoV M^{pro} (PDB Code 1uk4) diente als Referenzstruktur. Ein Datensatz von 3029 nicht-natürlicher Aminosäuren wurde durch eine Datenbanksuche in den Katalogen des Available Chemicals Directory (ACD) und von Sigma-Aldrich gewonnen[1]. Aus diesem Datensatz wurde mehrfach (unter verschiedenen Namen) abgelegte Aminosäuren und Aminosäuren mit einem Molekulargewicht größer als 800 Dalton herausgefiltert. Die verbleibende Bibliothek von 1230 nicht-natürlichen Aminosäuren wurde mit vorgegebenen Bindungsmodus der Peptidbindung durch Einschalten der MapRef Option in FlexX in die Bindetasche eingepaßt. Dazu wird die Geometrie des Basisfragmentes (hier die Peptidbindung (CA, N, C O) der jeweiligen Seitenkette aus dem Inhibitorpeptid) vorgegeben. Um die Verbindungen für das Docking vorzubereiten, wurden zunächst die Protonierungszustände angepasst. Alle exocyclischen Guanidino- and Amidinogruppen sowie primären und sekundären aliphatischen Aminogruppen wurden protoniert, während Phosphorsäure-, Sulfonsäure- und Carbonsäuregruppen deprotoniert wurden. Im Verlaufe der inkrementellen Aufbauphase wurde die FlexX-Scoringfunktion benutzt, die 10 besten Dockinglösungen wurden mit der FlexX-(Böhm)- und der Drugscore-Scoringfunktion nachbewertet. Tabelle 6.4 zeigt jeweils die für die drei untersuchten Subtaschen ausgewählten Ligandfragmente.

Der Scoring-Wert in Drugscore bewertet in der Summe alle Protein-Ligand-Kontakte im Bereich von 1 bis 6 Å [Gohlke et al., 2000a]. Die Bewertung eines Protein-Ligand-Komplexes ist somit abhängig von der Anzahl der Ligandatome und damit von seiner Größe. Eine graphische Auswertung der erhaltenen Dockingergebnisse hat solche Überlagerungen als positiv bewertet, die die jeweilige Subtasche gut ausfüllen, eine gute Komplementarität von Ligand und Protein zeigen und eine gute Bewertung nach Drugscore und/oder der Böhm-Scoringfunktion erhalten haben. Aminosäuren mit sehr großen Seitenketten wurden bei der graphischen Auswertung nicht berücksichtigt. In Tabelle 6.4 ist eine Auswahl der gefundenen Ligandfragmente vorgestellt, die nach visueller Beurteilung der Bausteine ausgewählt wurden.

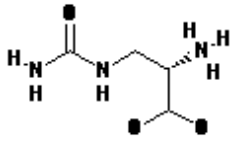
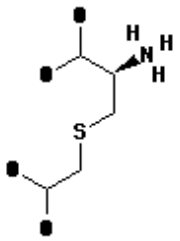
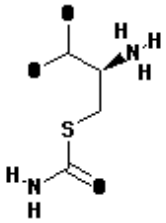
[1]Der Datensatz wurde von A. Evers und A. Hillebrecht zusammengestellt.

Tab. 6.4 Einige ausgewählte Aminosäuren aus der Docking-Studie. Dargestellt sind ausgewählte Aminosäuren, die in der Docking-Analyse gefunden wurden. Angegeben ist die SARS CoV M^{pro} Subtasche, die Aminosäure mit einer laufenden Nummer, die FlexX- und Drugscore-Bewertung und der jeweilige Rang in Klammern. Die Aminosäuren sind ohne polare Wasserstoffe und explizite Formalladungen dargestellt.

Tasche	Nummer	Verbindung	FlexX-Score (Rang)	DS-Score (Rang)
S1	**20**		-17.6 (2)	-5.96 (56)
S1	**21**		-16.98 (3)	-7.25 (8)
S1	**22**		-15.2 (8)	-7.21 (9)

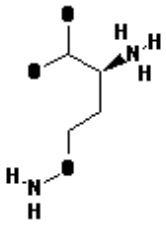
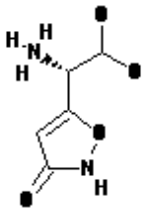
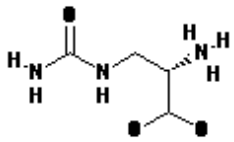
(Fortsetzung nächste Seite)

Fortsetzung Tab. 6.4

Tasche	Nummer	Verbindung	FlexX-Score (Rang)	DS-Score (Rang)
S1	**23**		-13.61 (17)	-4.17 (150)
S1	**24**		-12.83 (21)	-5.95 (58)
S1	**25**		-12.6 (22)	-5.16 (96)

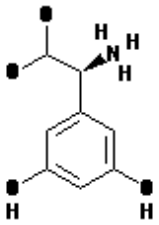
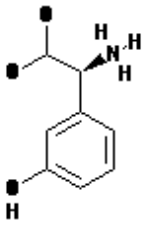
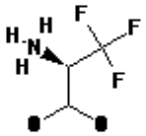
(Fortsetzung nächste Seite)

Fortsetzung Tab. 6.4

Tasche	Nummer	Verbindung	FlexX-Score (Rang)	DS-Score (Rang)
S1	**26**		-12.55 (23)	-5.42 (80)
S2	**27**		-2.38 (6)	-1.94 (69)
S2	**28**		-0.8 (11)	-2.18 (58)

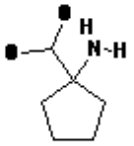
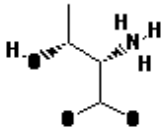
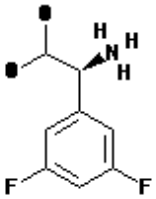
(Fortsetzung nächste Seite)

Fortsetzung Tab. 6.4

Tasche	Nummer	Verbindung	FlexX-Score (Rang)	DS-Score (Rang)
S2	**29**		-0.63 (12)	-3.90 (20)
S2	**30**		-0.23 (17)	-3.91 (19)
S2	**31**		0.0 (18)	-0.00 (122)

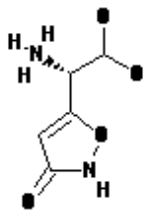
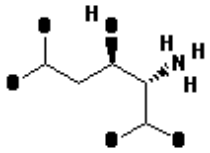
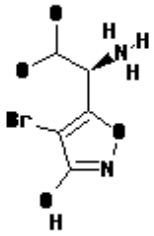
(Fortsetzung nächste Seite)

Fortsetzung Tab. 6.4

Tasche	Nummer	Verbindung	FlexX-Score (Rang)	DS-Score (Rang)
S2	**32**		0.0 (20)	-0.00 (124)
S2	**33**		0.6 (29)	-3.49 (30)
S2	**34**		0.8 (32)	4.76 (151)

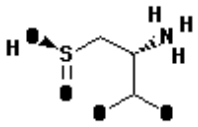
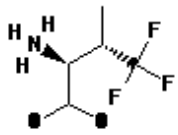
(Fortsetzung nächste Seite)

Fortsetzung Tab. 6.4

Tasche	Nummer	Verbindung	FlexX-Score (Rang)	DS-Score (Rang)
S4	**35**		-5.04 (3)	0.03 (132)
S4	**36**		-1.54 (20)	-3.17 (44)
S4	**37**		-1.53 (21)	-3.64 (29)

(Fortsetzung nächste Seite)

Fortsetzung Tab. 6.4

Tasche	Nummer	Verbindung	FlexX-Score (Rang)	DS-Score (Rang)
S4	**38**		0.95 (48)	-3.42 (34)
S4	**39**		3.45 (95)	-0.23 (127)

Um die Relevanz der erhaltenen Dockinglösungen abschätzen zu können, wurden die Aminosäuren, die die entsprechende Subtasche in der SARS CoV M^{pro} besetzen, mit eingepasst. Für die S1-Tasche erhält ein gedocktes Asparagin (in der Kristallstruktur befindet sich dort ein Glutamin) eine sehr gute Bewertung (FlexX-Score -15.0) und wird auf Rang 11 platziert. Die mit dem FlexX-Score berechneten Energiewerte für die S2-Tasche sind relativ hoch und die Wechselwirkung ist als ungünstig anzusehen. Die Größenordnung der Score-Werte liegt aber in dem Bereich, der für die Aminosäuren (Threonin oder Valin), die diese Stelle in den Kristallstrukturen besetzen, gefunden wird.

In einer Kooperation mit Prof. Dr. Rademann (FMP, Berlin-Buch) wird an der Entwicklung von peptidischen Inhibitoren für die SARS CoV M^{pro} gearbeitet. Durch die Cavbase und Docking Analysen konnten Vorschläge für peptidische Inhibitoren mit variierten Seitenketten getroffen werden (siehe Abbildung 6.4). Erste Verbindungen unter

Verwendung von natürlichen Aminosäuren wurden bereits synthetisiert, es liegen aber noch keine Affinitätsdaten vor.

6.6 Zusammenfassung und Schlussfolgerungen

In diesem Kapitel wurde der Einsatz von Cavbase bei der Ähnlichkeitssuche von Proteinsubtaschen dargestellt. Mit Hilfe von Cavbase gelingt eine physikochemische Charakterisierung von Subtaschen, die konsistent mit einer Bindetaschencharakterisierungen durch andere Methoden (*Hotspot*-Analyse, Dockingstudie) ist. Darüber hinaus liefert Cavbase konkrete Ideen über Ligandfragmente, die in ähnliche physikochemische Bindetaschenbereiche binden und erlaubt so das Auffinden von denkbaren bioisosteren Verbindungen. Mit zunehmender Zahl an gelösten Kristallstrukturen wird ein solcher Ansatz immer wertvollere Ideen für die Optimierung und das Design von Inhibitoren liefern. Zusätzlich konnten in einer Dockingstudie an der SARS CoV M^{pro} Vorschläge für die Synthese peptidartiger Inhibitoren getroffen werden.

7 Zusammenfassung und Ausblick

7.1 Zusammenfassung

Die Analyse der Ähnlichkeitsbeziehungen von Proteinbindetaschen ist das zentrale Thema, das in verschiedenen Variationen in dieser Arbeit bearbeitet wurde. Die grundlegende Annahme hinter dem vorgestellten Ansatz ist, daß die Funktion eines Proteins in starkem Maße von der Gestalt und den physikochemischen Eigenschaften der Bindetasche bestimmt wird. Ähnlichkeiten von Proteinbindetaschen lassen sich demnach dazu nutzen, Rückschlüsse über die Funktion von Proteinen zu erhalten. Außerdem helfen sie in der Identifizierung von Liganden oder Ligandfragmenten, die in verwandte Bindetaschen binden und können so Ideen für das Design von neuen Inhibitoren liefern. Des Weiteren können Ähnlichkeiten in den Bindetaschen nicht sequenz-verwandter Proteine dazu genutzt werden, mögliche Kreuzreaktivitäten zwischen diesen vorherzusagen. Schließlich bilden sie die Grundlage für eine funktionelle Klassifizierung von Proteinbindetaschen. In dieser Arbeit wurden deshalb Methoden entwickelt, um große Datensätzen von Bindetaschen miteinander vergleichen und Ähnlichkeits- und Clusteranalysen durchführen zu können. Die vorgestellte Arbeit baut auf der Methode Cavbase auf [Schmitt, 2000; Schmitt et al., 2002], die die geometrische Form und die physikochemischen Eigenschaften von Bindetaschen beschreibt und ähnliche Bereiche in zwei Bindetaschen bestimmt.

Eine schnelle und effiziente Durchführung von Bindetaschenvergleichen ist eine wichtige Voraussetzung für die Analyse von großen Datensätzen. Bindetaschenvergleiche konnten dadurch signifikant beschleunigt werden, indem heuristische Filter die Zahl der zeitlich aufwendigen Rechenschritte während der Überlagerung und der Bewertung von Cliquelösungen reduzierten. Das Vergleichsverfahren konnte um einen Faktor von 40 beschleunigt werden. Es ist nun möglich, Bindetaschenvergleiche mit sehr großen Datensätzen im Millionenmaßstab in einer realisierbaren Zeitdauer durchzuführen.

Die Analyse von großen Datensätzen verlangt einen effizienten Datenzugriff sowie eine effiziente Visualisierung. Die Ergebnisse eines Bindetaschenvergleichs werden in Form einer XML-Datei zurückgeliefert. Durch das Ablegen dieser Informationen in einer relationalen Datenbank (MySQL) und die Entwicklung von Suchanfragen konnten die

Ergebnisse der Ähnlichkeitsanalysen sehr gut ausgewertet werden. Zur Visualisierung wurde eine Schnittstelle zu dem Programm Pymol [DeLano, 2004] entwickelt, das einen interaktiven Umgang mit Bindetaschen erlaubt und die graphische Auswertung von Ähnlichkeiten in Bindetaschen stark vereinfacht.

Ein Vergleich der in Kristallstrukturen beobachteten Wechselwirkungen von Aminosäuren hat gezeigt, daß einige Interaktionsmöglichkeiten der Aminosäuren im bisherigen Ansatz durch Pseudozentren noch nicht wiedergegeben wurden (Abschnitt 3.2). Beispielsweise wurde die Fähigkeit der Seitenketten von Aspartat, Glutamat, Asparagin, Glutamin und Arginin zur Ausbildung von π-Wechselwirkungen nicht berücksichtigt. Deshalb wurde für diese Aminosäuren ein entsprechendes Pi Pseudozentrum eingeführt. Außerdem wurde die Fähigkeit der Thiolgruppe des Cysteins zur Ausbildung von Wasserstoffbrücken durch ein Donor Pseudozentrum berücksichtigt.

Pseudozentren exponieren ihre physikochemische Eigenschaft auf die Bindetaschenoberfläche. Eine Analyse von den Bereichen auf der Bindetaschenoberfläche, die von keiner Eigenschaft besetzt sind, hat gezeigt, daß die Seitenketten aromatischer Aminosäuren in Wechselwirkungsdistanz zu diesen Oberflächenbereichen liegen. Durch eine veränderte Repräsentation von aromatischen Seitenketten wird eine bessere Berücksichtigung der aromatischen Eigenschaften erreicht (Abschnitt 3.3.2).

Ein Vergleich der Bindetaschenrepräsentation mit wissensbasierten Ansätzen wie Superstar und Drugscore hat gezeigt (Abschnitt 3.4.1), daß die Eigenschaften der Bindetaschen von Cavbase sehr gut abgebildet werden. Im Fall der Analyse mit Drugscore konnte ein Verfahren für die Generierung und automatische Konturierung von *Hotspot*-Feldern etabliert werden. Damit ist es möglich, eine große Anzahl an Proteinen mit Drugscore zu untersuchen und die berechneten Drugscore *Hotspots* mit der Cavbase Bindetaschenoberfläche zu vergleichen.

Durch die Verwendung von Oberflächenpunkten, die mehrere physikochemische Eigenschaften gleichzeitig besitzen können, ist es möglich eine alternative Beschreibung der Bindetaschenoberfläche zu verwenden (Kapitel 3.3). Die Ergebnisse der Ähnlichkeitsanalyse unter Verwendung der Oberflächenpunkte mit multiplen Eigenschaften sind konsistent mit dem bisher verwendeten Verfahren. Es ist aber auch möglich, Ähnlichkeiten zu detektieren, die mit dem ursprünglichen Verfahren nicht gefunden wurden.

Oftmals ist es von besonderem Interesse, sich in der Vergleichsanalyse auf bestimmte Bereiche der Bindetasche zu konzentrieren. So zum Beispiel auf Bereiche, die für die

Funktion des Proteins eine wichtige Rolle spielen oder interessante Ligandfragmente gebunden haben. Durch die Verwendung von Proteinsubtaschen und dem Ablegen als eigene Objekte in Cavbase ist es möglich, effizient mit diesen Bereichen einer Bindetasche als Anfragetasche Ähnlichkeitssuchen durchzuführen. So konnten beispielsweise für die SARS CoV M^{pro} Protease Ähnlichkeitssuchen für die einzelnen Proteasesubtaschen durchgeführt und eine Charakterisierung der Bindetaschen vorgenommen werden (Kapitel 6). Diese Beschreibung ist konsistent mit Bindetaschencharakterisierungen durch andere Methoden (*Hotspot*-Analyse, Dockingstudien). Darüber hinaus liefert Cavbase konkrete Ideen über Ligandfragmente, die in ähnliche physikochemische Bindetaschenbereiche binden und erlaubt so das Auffinden von bioisosteren Verbindungen. Mit zunehmender Zahl an gelösten Kristallstrukturen wird ein solcher Ansatz immer wertvollere Ideen für die Optimierung und das Design von Inhibitoren liefern. Zusätzlich konnte in einer Dockingstudie an der SARS CoV M^{pro} Vorschläge für das Design von peptidartigen Inhibitoren gemacht werden.

Gerade in der Analyse von großen Datensätzen werden sehr viele Vergleiche zweier Bindetaschen berechnet, die keine große Ähnlichkeit zueinander zeigen. Deshalb wurde nach einem Verfahren gesucht, daß solche Fälle detektieren kann und dadurch die Anzahl der Berechnungen reduziert, die mit dem aufwendigeren Clique Vergleichsverfahren durchgeführt werden müssen. Dazu wurde das Konzept der Bitstringrepräsentation und -vergleiche auf Bindetaschen angewendet (Kapitel 3.6). Sie haben den Vorteil, daß sich die Ähnlichkeit zweier Bitstrings sehr schnell berechnen läßt. Dreiecke von Pseudozentren werden dazu benutzt, um die strukturelle Information und die physikochemischen Eigenschaften der Bindetasche in einem Bitstring zu kodieren. Vergleiche mit Bindetaschenbitstrings sind sehr gut in der Lage, unähnliche Bindetaschen aus einem Datensatz herauszufiltern. Das hat den Vorteil, daß die aufwendigere Cliquesuche für eine kleinere Anzahl an Bindetaschen durchgeführt werden muß.

An verschiedenen Beispielen ist der erfolgreiche Einsatz von Cavbase in der Ähnlichkeitsanalyse und dem Entdecken von verwandten Proteinbindetaschen gezeigt worden. So wurden Ähnlichkeitsanalysen für eine virale Cysteinprotease (Abschnitt 4.1.1), für ein Zink-bindendes Protein (Abschnitt 4.1.3) sowie für ein NAD(P)-bindendes Protein (Abschnitt 4.1.2) durchgeführt. Cavbase ist sicher in der Lage, zu den Suchanfragetaschen ähnliche Bindetaschen verwandter Proteine zu identifizieren. Gerade im letzten Fall konnten aber Ähnlichkeiten zu anderen Proteinen entdeckt werden, bei denen die entsprechenden Proteine keine Sequenz- und Faltungshomologie zeigen und die Bindung des Kofaktors über komplett verschiedene Aminosäuren vermittelt wird. Gerade solche

Ähnlichkeiten sind mit Verfahren, die auf Sequenz- oder Faltungshomologie beruhen, nicht zu entdecken. Weiterhin konnten an 26 ausgewählten Proteinen, deren Funktion zum Zeitpunkt der Kristallstrukturbestimmung noch nicht bekannt war (sogenannten *hypothetical proteins*), die Möglichkeiten und Grenzen des Cavbase Ansatzes gezeigt werden (Abschnitt 4.2). Cavbase ist in der Lage, funktionelle Ähnlichkeiten mit anderen Proteinen zu entdecken und Ideen für die Funktionsannotierung vorzuschlagen. Sind zu dem Protein unbekannter Funktion verwandte Proteine strukturell charakterisiert, dann ist es mit Cavbase möglich, diese ähnlichen Bereiche in den Bindetaschen zu detektieren. Eine besondere Herausforderung an die Funktionszuweisung ist die Annotation von Strukturen, für die nur wenig verwandte Strukturen bekannt sind. In diesen Fällen ist es möglich, mit Cavbase Informationen über Liganden oder Ligandfragmente zu sammeln, die in ähnliche Subtaschen binden und so Ideen über eine mögliche Funktion des Proteins zu bekommen.

Ähnlichkeiten in den Bindetaschen von zwei Proteinen können sich in einer unerwünschten Nebenwirkung manifestieren. Eine mögliche Kreuzreaktivität ist bei Sequenzverwandten Strukturen relativ einfach abzuschätzen. Schwieriger ist die Vorhersage von Kreuzreaktivitäten, wenn die betrachteten Proteine keine Sequenz- und Faltungsmusterhomologie aufweisen. In dieser Arbeit konnte eine im Experiment beobachtete Kreuzreaktivität strukturell verstanden und mit Cavbase nachvollzogen werden (siehe Kapitel 4.3). Cyclooxygenaseinhibitoren wie Celecoxib sind ebenfalls hochaffine Inhibitoren der Carboanhydrase II. Mit Cavbase konnten ähnliche Bereiche in beiden Bindetaschen detektiert werden. Übereinstimmungen zwischen beiden Bindetaschen wurde in diesem Fall besonders dadurch gefunden, indem man Subtaschen aus COX-2 zur Ähnlichkeitsanalyse benutzt und den Grad der dort entdeckten Ähnlichkeiten kombinierte. So war man in der Lage, Ähnlichkeiten zu finden, die man bei Vergleichen kompletter Bindetaschen nicht entdecken würde. Als zweiten Fall einer möglichen Kreuzreaktivtität wurden ähnliche Bereiche in den Bindetaschen von Carboanhydrase und Malatdehydrogenase aufgefunden, die eine im Experiment beobachtbare Kreuzreaktivität strukturell plausibel machen (siehe Kapitel 4.4.1). Den experimentellen Beweis hierfür muß eine Kristallstrukturbestimmung erbringen. Cavbase hat das Potential, Kreuzreaktivitäten zwischen nicht verwandten Proteinfamilien zu entdecken.

Traditionelle Methoden, die die Ähnlichkeit zwischen Proteinen oder Proteinbindetaschen bestimmen (Kapitel 2), vergleichen eine Bindetasche gegen einen großen Datensatz eben solcher und beschränken sich dabei nur auf die Analyse ausgewählter

Bindetaschen. In dieser Arbeit wurde der Fokus stattdessen auf eine Clusteranalyse von großen Datensätzen gelegt. Dabei steht jetzt nicht die Ähnlichkeitsbeziehungen einzelner Bindetaschen zueinander im Vordergrund, sondern die gleichzeitige Analyse von Ähnlichkeitsbeziehungen von mehreren Bindetaschen. Besonders interessant ist in diesem Zusammenhang die Analyse von Proteinfamilien. Am Beispiel von zwei pharmazeutisch relevanten Proteinfamilien den α-Carboanhydrasen und den Proteinkinasen konnte gezeigt werden, wie sich Ähnlichkeiten und Unterschiede in den Bindetaschen dazu nutzen lassen, um eine funktionelle Klassifizierung dieser Familien aufzubauen. Cavbase ist in der Lage, die betrachteten Bindetaschen auf der Ebene der Proteinsubfamilie zu unterscheiden. Im Fall der Carboanhydrasen wurden beispielsweise strukturelle Unterschiede innerhalb einer Subfamilie entdeckt, die wichtig für den katalytischen Mechanismus der Familie sind. Sogar im Fall der strukturell flexiblen Proteinfamilie der Proteinkinasen gelang eine sinnvolle Klassifizierung. Cavbase ist darüberhinaus in der Lage, zwischen Kinasen einer Proteinsubfamilie (beispielsweise CDK2, PKA), die in verschiedenen Aktivitätszuständen vorliegen, zu unterscheiden. Die erhaltenen Ergebnisse stehen im Einklang mit den Ergebnissen anderer Methoden, die die Proteine aufgrund Sequenz- und Faltungsmusterähnlichkeiten klassifizieren und dabei oft noch manuelle Annotationen einfließen lassen (z.B. SCOP oder CATH). Das ist umso erstaunlicher, da für Cavbase nur Wissen über die strukturelle Information der Bindetasche verwendet wird. Die Cavbase Klassifizierung unterscheidet sich von den anderen Verfahren vor allem dadurch, wie Beziehungen zwischen den einzelnen Proteinfamilien aufgebaut werden. So konnten eine Reihe von Kinasenpaare identifiziert werden, die unterschiedliche Verwandtschaften im Sequenz- und Cavbaseraum aufweisen(z.B. MAP $p38\alpha$ und MAP p38γ). Hier zeigt sich die Stärke des Cavbase Ansatzes, da Cavbase unabhängig von Sequenzhomologie Verwandtschaften in Proteinfamilien entdecken kann. Des Weiteren konnte die im Experiment beobachtete Kreuzreaktivität des Kinaseinhibitors Gleevec an c-Abl und c-Kit mit Cavbase strukturell nachvollzogen werden. Als am ähnlichsten erkannte Bindetasche zu c-Abl wurde die Bindetasche von c-Kit gefunden. Dies ist ein weiteres Beispiel für das Potential von Cavbase mit seiner Möglichkeit, Kreuzreaktivitäten aus der Struktur der Bindetasche ableiten und vorhersagen zu können.

7.2 Ausblick

In einer Validierungsstudie der Eigenschaftsrepräsentation der Aminosäuren in Cavbase konnten mit einem wissensbasierten Ansatz ausgezeichnete Wechselwirkungsfelder (*Hotspots*) in der Bindetasche berechnet werden. Diese *Hotspots* sollten sich ebenfalls zur kompletten Beschreibung von Proteinbindetaschen einsetzen lassen. Eine solche Beschreibung hätte neben der Konzentration auf Bereiche, die für die Protein-Ligand Wechselwirkung wichtig sind, auch den Vorteil, daß die Repräsentation der Bindetaschen durch weniger Koordinaten erfolgen könnte. Dies würde den Zeitbedarf in den Ähnlichkeitsanalysen stark verringern.

In einer Zusammenarbeit mit Prof. Hüllermeier und Nils Weskamp (Universität Marburg, Fachbereich Informatik) konnte eine weitere Beschleunigung des Vergleichsprozeßes durch den Einsatz eines kombinierten Clique-Hashing Verfahrens erreicht werden [Weskamp et al., 2004]. Dieser Ansatz verbindet die Vorteile einer Cliquedetektion mit *Geometric Hashing* Methoden. Die Vergleiche unter Verwendung der Oberflächen mit den multiplen Eigenschaften hat den Nachteil, daß sie zeitaufwendiger sind. Es bietet sich deshalb an, dieses Verfahren zur Ähnlichkeitssuche einzusetzen, wobei die physikochemische Modellierung der Bindetaschen auf den multiplen Oberflächenbereichen aufsetzt. Dadurch wird der Nachteil der längeren Laufzeit ausgeglichen.

In dieser Studie konnte am Beispiel der Proteinfamilien der Carboanhydrasen und Proteinkinasen sinnvolle Ergebnisse in der Klassifizierung von Proteinfamiliien erhalten werden. Die Klassifizierungsanalyse sollte noch für weitere Proteinfamilien durchgeführt werden. Aber es sind nicht nur Analysen mit kristallographisch aufgeklärten Strukturen denkbar, sondern es können auch Homologiemodelle von Proteinen für eine Klassifizierungsanalyse von Proteinbindetaschen benutzt werden. Initiale Analysen für die Familie der Carboanhydrasen haben erfolgversprechende Resultate gezeigt. Dabei stellt sich allerdings die Frage, ob die Qualität der generierten Homologiemodelle für Ähnlichkeitsanalysen in großen Datensätzen geeignet ist.

Die Klassifizierung von Proteinen, die mit Cavbase gewonnen werden, wurden mit existierenden Klassifizierungen basierend auf Sequenz- oder Faltungshomologie verglichen werden. In der Analyse von Ähnlichkeiten und Unterschieden zwischen beiden Klassifizierungsschemata können so wichtige Erkenntnisse über Kreuzbeziehungen zwischen Proteinen gewonnen werden. Weiterhin wäre es sehr interessant, die Cavbase Klassifi-

zierung mit den Affinitätsprofilen kleiner Molekülen zu validieren und zu vergleichen. Momentan sind solche Datensätze öffentlich nicht zugänglich und es bleibt zu hoffen, daß sich die Datenlage hier verbessert.

Durch die Zerlegung der Celecoxib-Bindetasche in COX-2 in einzelne Subtaschen und durch die anschließende Ähnlichkeitssuche mit ihnen konnte die Ähnlichkeit zwischen Carboanhydrasen und COX-2 entdeckt werden. Eine automatische Zerlegung aller Proteinbindetaschen in kleinere Subtaschen und eine Ähnlichkeitssuche mit den erzeugten Subtaschen sollte einige Fälle von bisher nicht bekannten Kreuzreaktivitäten systematisch erkennen lassen.

A Anhang

A.1 Verwendete Software und Hardware

Die Entwicklung erfolgte unter Debian *woody* in der Programmiersprache C++ mit dem GNU C/C++-Compiler in der Version 2.95.3 und 2.95.4 sowie in der Programmiersprache Python Version 2.3. Die Berechnungen der Bindetaschenvergleiche wurden auf einem Athlon 2100 Mhz und Pentium IV 3.2 GHz ausgeführt.

- Die im Rahmen dieser Arbeiten entwickelten Programme bauen auf Quellen der Rezeptor-Ligand Datenbank Relibase+ und deren Erweiterung Cavbase auf.
- Zur Berechnung der Superstar *Hotspots* wurde Superstar in der Version 1.4 verwendet.
- Zur Berechnung der Drugscore *Hotspots* wurde Drugscore 1.2 verwendet.
- Das Docking der Liganden wurde mit FlexX in der Version 1.13 durchgeführt.
- Zur Visualisierung der Bindetaschen wurde das Programm Pymol verwendet.

A.2 Bindetaschenvergleiche mit dem cliq5.lx Programm

A.2.1 Synopsis

```
./cliq5.lx [options] cavcode1 cavcode2 <-xml/-reli>
```

Das Programm `cliq5.lx` ist das Hauptprogramm, das zum Vergleich von Bindetaschen benutzt wird. Es entnimmt die Information über Bindetaschen entweder aus XML-Dateien oder direkt aus Relibase. Als Ausgabe wird eine XML-Datei mit Informationen über ähnlichen Pseudozentren und die Transformierung beider Taschen zurückgeliefert.

A.2.2 Verfügbare Optionen

Alle Optionen des Vergleichsprogramms sind über die Kommandozeile zugänglich und sind in Tabelle A.2.2 gelistet (Default-Parameter in eckigen Klammern).

Tab. A.1 Kommandozeilenoptionen für das cliq5.lx Programm.

Parameter	Wert	Kommentar
-c	Integers	listet die Kennnummern der Pseudozentren, die in der Ähnlichkeitsanalyse berücksichtigt werden sollen [aus]
-o	String	Verzeichnis, in dem die Ausgabedateien gespeichert werden [$PWD]
–xd	String	Verzeichnis, das die Bindetaschen XML-Dateien enthält [$RELIBASE_XML_DATA_DIR]
–log	Boolean	Eine Logdatei mit Namen pdbcav1_cav2.log wird angelegt [aus]
–simcenter_cliq	Boolean	Zum Aufbau des Clique Eingabegraphen wird Information über ähnliche Pseudozentren benutzt [an]
–simcenter_score	Boolean	Während des Scorings der Cliquelösungen wird Information über ähnliche Pseudozentren benutzt [ein]
–simsurf_cliq	Boolean	Zum Aufbau des Clique Eingabegraphen wird Information über ähnliche Oberflächenbereiche benutzt [aus]
–simsurf_score	Boolean	Während des Scorings der Cliquelösungen wird Information über ähnliche Oberflächenbereiche benutzt [aus]
–too_srfpair_orient	Boolean	Berücksichtigung der Ausrichtung der Pseudozentrenvektoren während des Aufbaus des Clique Eingabegraphs[aus]
–discardPeptide	Boolean	Pseudozentren aus der Peptidbindung werden nicht benutzt [aus]
–all_sco_sol	Boolean	Für alle Clique-Lösungen werden XML-Dateien als Ausgabe geschrieben [aus]
–onlyCliqueScore	Boolean	Berechnet die Bewertung der Clique-Lösungen nach $n\_clique \cdot \frac{1}{cli\_rms}$, die Oberflächenbeiträge werden nicht berücksichtigt [aus]
–signum_ctrs=n	Integer	wenn –onlyCliqueScore aktiviert ist, werden nur Lösungen mit mehr Pseudozentren berücksichtigt [aus]
–max_dist=n	Float	Maximale Distanz zwischen zwei Pseudozentren, die in der Ähnlichkeitsanalyse berücksichtigt werden [12Å]
–dist_tolerance=n	Float	Maximal erlaubte Distanz zwischen zwei Pseudozentren, um als ähnlich in der Ähnlichkeitsanalyse betrachtet zu werden [2Å]
–find_max_cliqsol=n	Integer	Anzahl an zurückgelieferten Clique-Lösungen [100]
–overlap_cutoff=n	Float	Zwei Oberflächenbereiche gelten als ähnlich, wenn sie zu mindestens [0.7] überlappen
–max_center_dist=n	Float	Maximale Distanz zweier Pseudozentren, um im Scoringschritt noch berücksichtigt zu werden [4.0Å]
–overlap_surfpatch=n	Float	Oberflächenbereiche müssen mehr als [0.7] überlappen
–max_surf_dist=n	Float	Maximale Distanz zwischen zwei Oberflächenpunkten, um im Scoring als ähnlich angesehen zu werden [1.0Å]

A.3 Visualisierung von Bindetaschen

Bindetaschen und Überlagerungen von zwei Bindetaschen werden in Pymol [DeLano, 2004] visualisiert. Dazu wurde eine Schnittstelle zwischen Cavbase und Pymol entwickelt. Informationen über Bindetaschen sowie Überlagerungen von zwei Bindetaschen können mit Pymol visualisiert werden.

```
$ alias pymol_cli 'pymol.com -xp ... '
```

Optionen zur Visualisierung
`$ readxml_cav.py - -file <cavity_file>`
Visualisierung einer Bindetaschen XML-Datei
`$ readxml_clisim.py - -file <clisim_file>`
Visualisierung eines Bindetaschenvergleichs, wobei Information über ähnliche Bereiche in Bindetaschen in einer XML-Datei abgelegt ist.
`$ readxml_clisim.py - -sql <db> <table> <cav1> <cav2>`
Visualisierung eines Bindetaschenvergleichs, wobei die Information über ähnliche Bereiche in Bindetaschen in einer MySQL Datenbank abgelegt ist.

Literaturverzeichnis

Sructural Proteomics In Europe, http://www.spineurope.org/.

F. Abbate, A. Casini, T. Owa, A. Scozzafava, and C. T. Supuran. Carbonic anhydrase inhibitors: E7070, a sulfonamide anticancer agent, potently inhibits cytosolic isozymes i and ii, and transmembrane, tumor-associated isozyme ix. *Bioorg Med Chem Lett*, 14(1):217–23, 2004.

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–10, 1990.

S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–402, 1997.

K. Anand, G. J. Palm, J. R. Mesters, S. G. Siddell, J. Ziebuhr, and R. Hilgenfeld. Structure of coronavirus main proteinase reveals combination of a chymotrypsin fold with an extra alpha-helical domain. *Embo J*, 21(13):3213–24, 2002.

K. Anand, J. Ziebuhr, P. Wadhwani, J. R. Mesters, and R. Hilgenfeld. Coronavirus main proteinase (3clpro) structure: basis for design of anti-sars drugs. *Science*, 300 (5626):1763–7, 2003.

V. Anantharaman, L. Aravind, and E. V. Koonin. Emergence of diverse biochemical activities in evolutionarily conserved structural scaffolds of proteins. *Curr Opin Chem Biol*, 7(1):12–20, 2003.

P. J. Artymiuk, A. R. Poirrette, H. M. Grindley, D. W. Rice, and P. Willett. A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. *J Mol Biol*, 243(2):327–44, 1994.

T. K. Attwood. The prints database: a resource for identification of protein families. *Brief Bioinform*, 3(3):252–63, 2002.

T. K. Attwood, P. Bradley, D. R. Flower, A. Gaulton, N. Maudling, A. L. Mitchell, G. Moulton, A. Nordle, K. Paine, P. Taylor, A. Uddin, and C. Zygouri. Prints and its automatic supplement, preprints. *Nucleic Acids Res*, 31(1):400–2, 2003.

O. Bachar, D. Fischer, R. Nussinov, and H. Wolfson. A computer vision based technique for 3-d sequence-independent structural comparison of proteins. *Protein Eng*, 6(3): 279–88., 1993.

A. Bairoch. The enzyme database in 2000. *Nucleic Acids Res*, 28(1):304–5, 2000.

G. J. Bartlett, C. T. Porter, N. Borkakoti, and J. M. Thornton. Analysis of catalytic residues in enzyme active sites. *J Mol Biol*, 324(1):105–21, 2002.

G. J. Bartlett, A. E. Todd, and J. M. Thornton. Inferring protein function from structure. *Methods Biochem Anal*, 44:387–407, 2003.

A. Bateman, E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-Jones, K. L. Howe, M. Marshall, and E. L. Sonnhammer. The pfam protein families database. *Nucleic Acids Res*, 30(1):276–80, 2002.

R. Batra, D. Christendat, A. Edwards, C. Arrowsmith, and L. Tong. Crystal structure of mth169, a crucial component of phosphoribosylformylglycinamidine synthetase. *Proteins*, 49(2):285–8, 2002.

S. Bellon, M. J. Fitzgibbon, T. Fox, H. M. Hsiao, and K. P. Wilson. The structure of phosphorylated p38gamma is monomeric and reveals a conserved activation-loop conformation. *Structure Fold Des*, 7(9):1057–65, 1999.

D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. Genbank. *Nucleic Acids Res*, 31(1):23–7, 2003.

H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Res*, 28(1): 235–42., 2000.

T. Blundell, H. Jhoti, and C. Abell. High-throughput crystallography for lead discovery in drug design. *Nat.Rev.Drug Discov.*, 1(1):45–54, 2002.

B. Boeckmann, A. Bairoch, R. Apweiler, M. C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider. The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic Acids Res*, 31(1):365–70, 2003.

C. Branden and J. Tooze. *Introduction to protein structure*. Garland, 1999.

J. A. Brannigan, G. Dodson, H. J. Duggleby, P. C. Moody, J. L. Smith, D. R. Tomchick, and A. G. Murzin. A protein catalytic framework with an n-terminal nucleophile is capable of self-activation. *Nature*, 378(6555):416–9, 1995.

S. E. Brenner and M. Levitt. Expectations from structural genomics. *Protein Sci*, 9 (1):197–200, 2000.

C. Bron and J. Kerbosch. Algorithm 457. finding all cliques of an undirected graph. *Commun. ACM*, 16(9):575–577, 1973.

I. J. Bruno, J. C. Cole, J. P. Lommerse, R. S. Rowland, R. Taylor, and M. L. Verdonk. Isostar: a library of information about nonbonded interactions. *J Comput Aided Mol Des*, 11(6):525–37, 1997.

P. Bucher and A. Bairoch. A generalized profile syntax for biomolecular sequence motifs and its function in automatic sequence interpretation. *Proc Int Conf Intell Syst Mol Biol*, 2:53–61, 1994.

S. K. Burley. An overview of structural genomics. *Nat Struct Biol*, 7 Suppl:932–4, 2000.

S. K. Burley and J. B. Bonanno. Structural genomics. *Methods Biochem Anal*, 44: 591–612, 2003.

B. J. Canagarajah, A. Khokhlatchev, M. H. Cobb, and E. J. Goldsmith. Activation mechanism of the map kinase erk2 by dual phosphorylation. *Cell*, 90(5):859–69, 1997.

R. Capdeville, E. Buchdunger, J. Zimmermann, and A. Matter. Glivec (sti571, imatinib), a rationally developed, targeted anticancer drug. *Nat.Rev.Drug Discov.*, 1(7): 493–502, 2002.

O. Carugo and S. Pongor. Recent progress in protein 3d structure comparison. *Curr Protein Pept Sci*, 3(4):441–9, 2002.

A. Casini, A. Scozzafava, A. Mastrolorenzo, and L. T. Supuran. Sulfonamides and sulfonylated derivatives as anticancer agents. *Curr Cancer Drug Targets*, 2(1):55–75, 2002.

C. I. Chang, B. E. Xu, R. Akella, M. H. Cobb, and E. J. Goldsmith. Crystal structures of map kinase p38 complexed to the docking sites on its nuclear substrate mef2a and activator mkk3b. *Mol Cell*, 9(6):1241–9, 2002.

A. D. Chapman, A. Cortes, T. R. Dafforn, A. R. Clarke, and R. L. Brady. Structural basis of substrate specificity in malate dehydrogenases: crystal structure of a ternary complex of porcine cytoplasmic malate dehydrogenase, alpha-ketomalonate and tetrahydonad. *J Mol Biol*, 285(2):703–12, 1999.

S. Cheek, H. Zhang, and N. V. Grishin. Sequence and structure classification of kinases. *J Mol Biol*, 320(4):855–81, 2002.

M. Cherry and D. H. Williams. Recent kinase and kinase inhibitor x-ray structures: mechanisms of inhibition and selectivity insights. *Curr Med Chem*, 11(6):663–73, 2004.

C. Chothia and A. M. Lesk. The relation between the divergence of sequence and structure in proteins. *Embo J*, 5(4):823–6, 1986.

D. Christendat, V. Saridakis, Y. Kim, P. A. Kumar, X. Xu, A. Semesi, A. Joachimiak, C. H. Arrowsmith, and A. M. Edwards. The crystal structure of hypothetical protein mth1491 from methanobacterium thermoautotrophicum. *Protein Sci*, 11(6):1409–14, 2002.

D. Christendat, A. Yee, A. Dharamsi, Y. Kluger, M. Gerstein, C. H. Arrowsmith, and A. M. Edwards. Structural proteomics: prospects for high throughput sample preparation. *Prog Biophys Mol Biol*, 73(5):339–45, 2000.

P. Cohen. Protein kinases–the major drug targets of the twenty-first century? *Nat Rev Drug Discov*, 1(4):309–15, 2002.

I. H. G. S. Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004.

A. Cuenda, J. Rouse, Y. Doza, R. Meier, P. Cohen, T. Gallagher, P. Young, and J. Lee. Sb 203580 is a specific inhibitor of a map kinase homologue which is stimulated by cellular stresses and interleukin-1. *FEBS Lett.*, 364(2):229–233, 1995.

J. Dancey and E. A. Sausville. Issues and progress with protein kinase inhibitors for cancer treatment. *Nat Rev Drug Discov*, 2(4):296–313, 2003.

K. Das, R. Xiao, E. Wahlberg, F. Hsu, C. H. Arrowsmith, G. T. Montelione, and E. Arnold. X-ray crystal structure of mth938 from methanobacterium thermoautotrophicum at 2.2 a resolution reveals a novel tertiary protein fold. *Proteins*, 45(4): 486–8, 2001.

M. de Rinaldis, G. Ausiello, G. Cesareni, and M. Helmer-Citterich. Three-dimensional profiles: a new tool to identify protein surface similarities. *J Mol Biol*, 284(4):1211–21., 1998.

W. L. DeLano. The pymol molecular graphics system., 2004. DeLano Scientific LLC, San Carlos, CA, USA, http://www.pymol.org.

S. Dietmann, J. Park, C. Notredame, A. Heger, M. Lappe, and L. Holm. A fully automatic evolutionary classification of protein folds: Dali domain dictionary version 3. *Nucleic Acids Res*, 29(1):55–7., 2001.

C. Dittrich, H. Dumez, H. Calvert, A. Hanauske, M. Faber, J. Wanders, M. Yule, M. Ravic, and P. Fumoleau. Phase i and pharmacokinetic study of e7070, a chloroindolylsulfonamide anticancer agent, administered on a weekly schedule to patients with solid tumors. *Clin Cancer Res*, 9(14):5195–204, 2003.

H. Doong, A. Vrailas, and E. C. Kohn. What's in the 'bag'?–a functional domain analysis of the bag-family proteins. *Cancer Lett*, 188(1-2):25–32, 2002.

G. M. Downs and P. Willett. Similarity searching in databases of chemical structures. In K. B. Lipkowitz and D. B. Boyd, editors, *Reviews in Computational Chemistry*, volume 7, pages 1–66. VCH Publishers, New York, 1996.

P. S. Dragovich, T. J. Prins, R. Zhou, S. E. Webber, J. T. Marakovits, S. A. Fuhrman, A. K. Patick, D. A. Matthews, C. A. Lee, C. E. Ford, B. J. Burke, P. A. Rejto, T. F. Hendrickson, T. Tuntland, E. L. Brown, r. Meador, J. W., R. A. Ferre, J. E. Harr, M. B. Kosa, and S. T. Worland. Structure-based design, synthesis, and biological evaluation of irreversible human rhinovirus 3c protease inhibitors. 4. incorporation of p1 lactam moieties as l-glutamine replacements. *J Med Chem*, 42(7):1213–24, 1999.

J. Drews. Drug discovery today - and tomorrow. *Drug Discov Today*, 5(1):2–4, 2000.

B. Druker. David a. karnofsky award lecture. imatinib as a paradigm of targeted therapies. *J.Clin.Oncol.*, 21(23 Suppl):239s–245s, 2003.

S. R. Eddy. Hidden markov models. *Curr Opin Struct Biol*, 6(3):361–5, 1996.

I. Eidhammer, I. Jonassen, and W. R. Taylor. Structure comparison and structure patterns. *J Comput Biol*, 7(5):685–716, 2000.

E. Eisenstein, G. L. Gilliland, O. Herzberg, J. Moult, J. Orban, R. J. Poljak, L. Banerjei, D. Richardson, and A. J. Howard. Biological function made crystal clear - annotation of hypothetical proteins via structural genomics. *Curr Opin Biotechnol*, 11(1):25–30, 2000.

R. A. Engh and D. Bossemeyer. Structural aspects of protein kinase control-role of conformational flexibility. *Pharmacol Ther*, 93(2-3):99–111, 2002.

L. Falquet, M. Pagni, P. Bucher, N. Hulo, C. J. Sigrist, K. Hofmann, and A. Bairoch. The prosite database, its status in 2002. *Nucleic Acids Res*, 30(1):235–8, 2002.

D. Fischer, S. L. Lin, H. L. Wolfson, and R. Nussinov. A geometry-based suite of molecular docking processes. *J Mol Biol*, 248(2):459–77., 1995a.

D. Fischer, R. Norel, H. Wolfson, and R. Nussinov. Surface motifs by a computer vision technique: searches, detection, and implications for protein-ligand recognition. *Proteins*, 16(3):278–92., 1993b.

D. Fischer, C. J. Tsai, R. Nussinov, and H. Wolfson. A 3d sequence-independent representation of the protein data bank. *Protein Eng*, 8(10):981–97., 1995b.

D. Fischer, H. Wolfson, S. L. Lin, and R. Nussinov. Three-dimensional, sequence order-independent structural comparison of a serine protease against the crystallographic database reveals active site similarities: potential implications to evolution and to protein folding. *Protein Sci*, 3(5):769–78., 1994.

D. Fischer, H. Wolfson, and R. Nussinov. Spatial, sequence-order-independent structural comparison of alpha/beta proteins: evolutionary implications. *J Biomol Struct Dyn*, 11(2):367–80., 1993a.

C. E. Fitzgerald, S. B. Patel, J. W. Becker, P. M. Cameron, D. Zaller, V. B. Pikounis, S. J. O'Keefe, and G. Scapin. Structural basis for p38alpha map kinase quinazolinone and pyridol-pyrimidine inhibitor specificity. *Nat Struct Biol*, 10(9):764–9, 2003.

G. A. FitzGerald and C. Patrono. The coxibs, selective inhibitors of cyclooxygenase-2. *N Engl J Med*, 345(6):433–42, 2001.

R. A. Fouchier, T. Kuiken, M. Schutten, G. van Amerongen, G. J. van Doornum, B. G. van den Hoogen, M. Peiris, W. Lim, K. Stohr, and A. D. Osterhaus. Aetiology: Koch's postulates fulfilled for sars virus. *Nature*, 423(6937):240, 2003.

T. Fox, J. T. Coll, X. Xie, P. J. Ford, U. A. Germann, M. D. Porter, S. Pazhanisamy, M. A. Fleming, V. Galullo, M. S. Su, and K. P. Wilson. A single amino acid substitution makes erk2 susceptible to pyridinyl imidazole inhibitors of p38 map kinase. *Protein Sci*, 7(11):2249–55, 1998.

F. Frye. Structure-activity relationship homology (sarah): a conceptual framework for drug discovery in the genomic era. *Chem. Biol.*, 6(1):R3–7, 1999.

H. H. Gan, R. A. Perlow, S. Roy, J. Ko, M. Wu, J. Huang, S. Yan, A. Nicoletta, J. Vafai, D. Sun, L. Wang, J. E. Noah, S. Pasquali, and T. Schlick. Analysis of protein sequence/structure similarity relationships. *Biophys J*, 83(5):2781–91, 2002.

J. Gasteiger and T. Engel, editors. *Chemoinformatics. A Textbook*. Wiley-VCH, Weinheim, first edition, 2003.

J. F. Gibrat, T. Madej, and S. H. Bryant. Surprising similarities in structure comparison. *Curr Opin Struct Biol*, 6(3):377–85, 1996.

W. B. Gleason, Z. Fu, J. Birktoft, and L. Banaszak. Refined crystal structure of mitochondrial malate dehydrogenase from porcine heart and the consensus structure for dicarboxylic acid oxidoreductases. *Biochemistry*, 33(8):2078–88, 1994.

J. Godden, L. Xue, F. Stahura, and J. Bajorath. Searching for molecules with similar biological activity: analysis by fingerprint profiling. *Pac.Symp.Biocomput.*, pages 566–575, 2000.

A. Godzik. The structural alignment between two proteins: is there a unique answer? *Protein Sci*, 5(7):1325–38, 1996.

H. Gohlke, M. Hendlich, and G. Klebe. Knowledge-based scoring function to predict protein-ligand interactions. *J Mol Biol*, 295(2):337–56, 2000a.

H. Gohlke and G. Klebe. Statistical potentials and scoring functions applied to protein-ligand binding. *Curr Opin Struct Biol*, 11(2):231–5, 2001.

H. Gohlke, H. M., and K. G. Predicting binding modes, binding affinities and 'hot spots' for protein-ligand complexes using a knowledge-based scoring function. *Persp Drug Design Discov*, 20:115–144, 2000b.

S. Goldsmith-Fischman and B. Honig. Structural genomics: computational methods for structure analysis. *Protein Sci.*, 12(9):1813–1821, 2003.

A. C. Good, T. J. Ewing, D. A. Gschwend, and I. D. Kuntz. New molecular shape descriptors: application in database screening. *J Comput Aided Mol Des*, 9(1):1–12, 1995.

A. C. Good and I. D. Kuntz. Investigating the extension of pairwise distance pharmacophore measures to triplet-based descriptors. *J Comput Aided Mol Des*, 9(4):373–9, 1995.

P. J. Goodford. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J Med Chem*, 28(7):849–57, 1985.

C. R. Goward and D. J. Nicholls. Malate dehydrogenase: a model for structure, evolution, and catalysis. *Protein Sci*, 3(10):1883–8, 1994.

M. Gribskov, R. Luthy, and D. Eisenberg. Profile analysis. *Methods Enzymol*, 183: 146–59, 1990.

M. Gribskov, A. D. McLachlan, and D. Eisenberg. Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci U S A*, 84(13):4355–8, 1987.

H. M. Grindley, P. J. Artymiuk, D. W. Rice, and P. Willett. Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *J Mol Biol*, 229(3):707–21, 1993.

R. J. Gum, M. M. McLaughlin, S. Kumar, Z. Wang, M. J. Bower, J. C. Lee, J. L. Adams, G. P. Livi, E. J. Goldsmith, and P. R. Young. Acquisition of sensitivity of stress-activated protein kinases to the p38 inhibitor, sb 203580, by alteration of one or more amino acids within the atp binding pocket. *J Biol Chem*, 273(25):15605–10, 1998.

J. Gunther, A. Bergner, M. Hendlich, and G. Klebe. Utilising structural knowledge in drug design strategies: applications using relibase. *J Mol Biol*, 326(2):621–36, 2003.

T. Hamelryck. Efficient identification of side-chain patterns using a multidimensional index tree. *Proteins*, 51(1):96–108, 2003.

M. Hendlich. Databases for protein-ligand complexes. *Acta Crystallogr D Biol Crystallogr*, 54(Pt 6 Pt 1):1178–82, 1998.

M. Hendlich, A. Bergner, J. Gunther, and G. Klebe. Relibase: design and development of a database for comprehensive analysis of protein-ligand interactions. *J Mol Biol*, 326(2):607–20, 2003.

M. Hendlich, F. Rippmann, and G. Barnickel. Ligsite: automatic and efficient detection of potential small molecule-binding sites in proteins. *J Mol Graph Model*, 15(6):359–63, 389, 1997.

J. G. Henikoff, E. A. Greene, S. Pietrokovski, and S. Henikoff. Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res*, 28(1):228–30, 2000.

J. G. Henikoff and S. Henikoff. Blocks database and its applications. *Methods Enzymol*, 266:88–105, 1996.

S. Henikoff, J. G. Henikoff, and S. Pietrokovski. Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics*, 15 (6):471–9, 1999.

L. Holm and C. Sander. Searching protein structure databases has come of age. *Proteins*, 19(3):165–73, 1994.

L. Holm and C. Sander. The fssp database: fold classification based on structure-structure alignment of proteins. *Nucleic Acids Res*, 24(1):206–9., 1996a.

L. Holm and C. Sander. Mapping the protein universe. *Science*, 273(5275):595–603, 1996b.

L. Holm and C. Sander. Touring protein fold space with dali/fssp. *Nucleic Acids Res*, 26(1):316–9, 1998.

A. Hopkins and C. Groom. The druggable genome. *Nat.Rev.Drug Discov.*, 1(9):727–730, 2002.

M. Huse and J. Kuriyan. The conformational plasticity of protein kinases. *Cell*, 109 (3):275–82, 2002.

E. Jacoby, A. Schuffenhauer, and P. Floersheim. Chemogenomics knowledge-based strategies in drug discovery. *Drug News Perspect.*, 16(2):93–102, 2003.

A. Jain, M. Murty, and P. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, 1999.

M. Jambon, A. Imberty, G. Deleage, and C. Geourjon. A new bioinformatic approach to detect common 3d sites in protein structures. *Proteins*, 52(2):137–45, 2003.

G. L. Johnson and R. Lapadat. Mitogen-activated protein kinase pathways mediated by erk, jnk, and p38 protein kinases. *Science*, 298(5600):1911–2, 2002.

G. Karypis. Cluto - a clustering toolkit - release 2.1, 2002.

M. A. Kastenholz, M. Pastor, G. Cruciani, E. E. Haaksma, and T. Fox. Grid/cpca: a new computational tool to design selective ligands. *J Med Chem*, 43(16):3033–44, 2000.

J. P. Keller, P. M. Smith, J. Benach, D. Christendat, G. T. deTitta, and J. F. Hunt. The crystal structure of mt0146/cbit suggests that the putative precorrin-8w decarboxylase is a methyltransferase. *Structure (Camb)*, 10(11):1475–87, 2002.

C. Y. Kim, D. A. Whittington, J. S. Chang, J. Liao, J. A. May, and D. W. Christianson. Structural aspects of isozyme selectivity in the binding of inhibitors to carbonic anhydrases ii and iv. *J Med Chem*, 45(4):888–93, 2002.

K. Kinoshita, J. Furui, and H. Nakamura. Identification of protein functions from a molecular surface database, ef-site. *J Struct Funct Genomics*, 2(1):9–22, 2002.

K. Kinoshita and H. Nakamura. Identification of protein biochemical functions by similarity search using the molecular surface database ef-site. *Protein Sci*, 12(8): 1589–95, 2003.

G. J. Kleywegt. Recognition of spatial motifs in protein structures. *J Mol Biol*, 285 (4):1887–97, 1999.

G. J. Kleywegt, T. Bergfors, H. Senn, P. Le Motte, B. Gsell, K. Shudo, and T. A. Jones. Crystal structures of cellular retinoic acid binding proteins i and ii in complex with all-trans-retinoic acid and a synthetic retinoid. *Structure*, 2(12):1241–58, 1994.

P. Koehl. Protein structure similarities. *Curr Opin Struct Biol*, 11(3):348–53, 2001.

A. Krogh, M. Brown, I. S. Mian, K. Sjolander, and D. Haussler. Hidden markov models in computational biology. applications to protein modeling. *J Mol Biol*, 235(5):1501–31, 1994.

T. Kuiken, R. A. Fouchier, M. Schutten, G. F. Rimmelzwaan, G. van Amerongen, D. van Riel, J. D. Laman, T. de Jong, G. van Doornum, W. Lim, A. E. Ling, P. K.

Chan, J. S. Tam, M. C. Zambon, R. Gopal, C. Drosten, S. van der Werf, N. Escriou, J. C. Manuguerra, K. Stohr, J. S. Peiris, and A. D. Osterhaus. Newly discovered coronavirus as the primary cause of severe acute respiratory syndrome. *Lancet*, 362 (9380):263–70, 2003.

S. Kumar, J. Boehm, and J. C. Lee. p38 map kinases: key signalling molecules as therapeutic targets for inflammatory diseases. *Nat Rev Drug Discov*, 2(9):717–26, 2003.

E. S. Lander et al. Initial sequencing and analysis of the human genome. *Nature*, 409 (6822):860–921., 2001.

R. A. Laskowski, J. D. Watson, and J. M. Thornton. From protein structure to biochemical function? *J Struct Funct Genomics*, 4(2-3):167–77, 2003.

A. R. Leach. *Molecular Modelling - Principles and Applications*. Pearson Education Limited, Harlow, Essex, second edition, 2001.

J. C. Lee, S. Kassis, S. Kumar, A. Badger, and J. L. Adams. p38 mitogen-activated protein kinase inhibitors–mechanisms and therapeutic potentials. *Pharmacol Ther*, 82(2-3):389–97, 1999.

J. C. Lee, J. T. Laydon, P. C. McDonnell, T. F. Gallagher, S. Kumar, D. Green, D. McNulty, M. J. Blumenthal, J. R. Heys, S. W. Landvatter, and et al. A protein kinase involved in the regulation of inflammatory cytokine biosynthesis. *Nature*, 372 (6508):739–46, 1994.

J. V. Lehtonen, K. Denessiouk, A. C. May, and M. S. Johnson. Finding local structural similarities among families of unrelated protein structures: a generic non-linear alignment algorithm. *Proteins*, 34(3):341–55., 1999.

M. P. Liang, D. R. Banatao, T. E. Klein, D. L. Brutlag, and R. B. Altman. Webfeature: An interactive web tool for identifying and visualizing functional sites on macromolecular structures. *Nucleic Acids Res*, 31(13):3324–7, 2003b.

S. L. Lin and R. Nussinov. Molecular recognition via face center representation of a molecular surface. *J Mol Graph*, 14(2):78–90, 95–7., 1996.

S. L. Lin, R. Nussinov, D. Fischer, and H. J. Wolfson. Molecular surface representations by sparse critical points. *Proteins*, 18(1):94–101, 1994.

S. Lindskog. Structure and mechanism of carbonic anhydrase. *Pharmacol Ther*, 74(1): 1–20, 1997.

L. Lo Conte, S. E. Brenner, T. J. Hubbard, C. Chothia, and A. G. Murzin. Scop database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res*, 30(1):264–7, 2002.

T. Lugo, A. Pendergast, A. Muller, and O. Witte. Tyrosine kinase activity and transformation potency of bcr-abl oncogene products. *Science*, 247(4946):1079–1082, 1990.

C. D. Mackereth, C. H. Arrowsmith, A. M. Edwards, and L. P. McIntosh. Zinc-bundle structure of the essential rna polymerase subunit rpb10 from methanobacterium thermoautotrophicum. *Proc Natl Acad Sci U S A*, 97(12):6316–21, 2000.

G. Manning, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam. The protein kinase complement of the human genome. *Science*, 298(5600):1912–34, 2002.

A. C. Martin, C. A. Orengo, E. G. Hutchinson, S. Jones, M. Karmirantzou, R. A. Laskowski, J. B. Mitchell, C. Taroni, and J. M. Thornton. Protein folds and functions. *Structure*, 6(7):875–84., 1998.

A. Matter. Tumor angiogenesis as a therapeutic target. *Drug Discov.Today*, 6(19): 1005–1024, 2001.

D. A. Matthews, P. S. Dragovich, S. E. Webber, S. A. Fuhrman, A. K. Patick, L. S. Zalman, T. F. Hendrickson, R. A. Love, T. J. Prins, J. T. Marakovits, R. Zhou, J. Tikhe, C. E. Ford, J. W. Meador, R. A. Ferre, E. L. Brown, S. L. Binford, M. A. Brothers, D. M. DeLisle, and S. T. Worland. Structure-assisted design of mechanism-based irreversible inhibitors of human rhinovirus 3c protease with potent antiviral activity against multiple rhinovirus serotypes. *Proc Natl Acad Sci U S A*, 96(20): 11000–7, 1999.

I. K. McDonald and J. M. Thornton. Satisfying hydrogen bonding potential in proteins. *J Mol Biol*, 238(5):777–93, 1994.

MCSG. Midwest center for structural genomics, http://www.mcsg.anl.gov/.

E. Meyer, R. Castellano, and F. Diederich. Interactions with aromatic rings in chemical and biological recognition. *Angew.Chem.Int.Ed Engl.*, 42(11):1210–1250, 2003.

J. B. Mitchell, C. L. Nandi, I. K. McDonald, J. M. Thornton, and S. L. Price. Amino/aromatic interactions in proteins: is the evidence stacked against hydrogen bonding? *J Mol Biol*, 239(2):315–31, 1994.

P. R. Mittl and M. G. Grutter. Structural genomics: opportunities and challenges. *Curr Opin Chem Biol*, 5(4):402–8, 2001.

A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. Scop: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247(4):536–40, 1995.

N. Nagano, C. A. Orengo, and J. M. Thornton. One fold with many functions: the evolutionary relationships between tim barrel families based on their sequences, structures and functions. *J Mol Biol*, 321(5):741–65, 2002.

B. Nagar, W. G. Bornmann, P. Pellicena, T. Schindler, D. R. Veach, W. T. Miller, B. Clarkson, and J. Kuriyan. Crystal structures of the kinase domain of c-abl in complex with the small molecule inhibitors pd173955 and imatinib (sti-571). *Cancer Res*, 62(15):4236–43, 2002.

T. Naumann and H. Matter. Structural classification of protein kinases using 3d molecular interaction field analysis of their ligand binding sites: target family landscapes. *J Med Chem*, 45(12):2366–78, 2002.

S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48(3):443–53, 1970.

J. Nissink, C. Murray, M. Hartshorn, M. Verdonk, J. Cole, and R. Taylor. A new test set for validating predictions of protein-ligand interaction. *Proteins*, 49(4):457–471, 2002.

M. E. Noble, J. A. Endicott, and L. N. Johnson. Protein kinase inhibitors: insights into drug design from structure. *Science*, 303(5665):1800–5, 2004.

B. Nolen, S. Taylor, and G. Ghosh. Regulation of protein kinases; controlling activity through activation segment conformation. *Mol.Cell*, 15(5):661–675, 2004.

R. Norel, S. L. Lin, H. J. Wolfson, and R. Nussinov. Shape complementarity at protein-protein interfaces. *Biopolymers*, 34(7):933–40., 1994.

Y. Oda, T. Owa, T. Sato, B. Boucher, S. Daniels, H. Yamanaka, Y. Shinohara, A. Yokoi, J. Kuromitsu, and T. Nagasu. Quantitative chemical proteomics for identifying candidate drug targets. *Anal Chem*, 75(9):2159–65, 2003.

C. Oefner, A. D'Arcy, M. Hennig, F. K. Winkler, and G. E. Dale. Structure of human neutral endopeptidase (neprilysin) complexed with phosphoramidon. *J Mol Biol*, 296 (2):341–9, 2000.

C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton. Cath–a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–108., 1997.

C. A. Orengo, I. Sillitoe, G. Reeves, and F. M. Pearl. Review: what can structural classifications reveal about protein evolution? *J Struct Biol*, 134(2-3):145–65., 2001.

C. A. Orengo, A. E. Todd, and J. M. Thornton. From protein structure to function. *Curr Opin Struct Biol*, 9(3):374–82., 1999.

T. Owa, H. Yoshino, T. Okauchi, K. Yoshimatsu, Y. Ozawa, N. H. Sugi, T. Nagasu, N. Koyanagi, and K. Kitoh. Discovery of novel antitumor sulfonamides targeting g1 phase of the cell cycle. *J Med Chem*, 42(19):3789–99, 1999.

Y. Ozawa, N. H. Sugi, T. Nagasu, T. Owa, T. Watanabe, N. Koyanagi, H. Yoshino, K. Kitoh, and K. Yoshimatsu. E7070, a novel sulphonamide agent with potent antitumour activity in vitro and in vivo. *Eur J Cancer*, 37(17):2275–82, 2001.

C. Pargellis, L. Tong, L. Churchill, P. F. Cirillo, T. Gilmore, A. G. Graham, P. M. Grob, E. R. Hickey, N. Moss, S. Pav, and J. Regan. Inhibition of p38 map kinase by utilizing a novel allosteric binding site. *Nat Struct Biol*, 9(4):268–72, 2002.

J. Park, K. Karplus, C. Barrett, R. Hughey, D. Haussler, T. Hubbard, and C. Chothia. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J Mol Biol*, 284(4):1201–10, 1998.

S. B. Patel, P. M. Cameron, B. Frantz-Wattley, E. O'Neill, J. W. Becker, and G. Scapin. Lattice stabilization and enhanced diffraction in human p38 alpha crystals by protein engineering. *Biochim Biophys Acta*, 1696(1):67–73, 2004.

F. M. Pearl, D. Lee, J. E. Bray, I. Sillitoe, A. E. Todd, A. P. Harrison, J. M. Thornton, and C. A. Orengo. Assigning genomic sequences to cath. *Nucleic Acids Res*, 28(1): 277–82, 2000.

W. R. Pearson. Rapid and sensitive sequence comparison with fastp and fasta. *Methods Enzymol*, 183:63–98, 1990.

W. R. Pearson and D. J. Lipman. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, 85(8):2444–8, 1988.

M. Pellecchia, D. Sem, and K. Wuthrich. Nmr in drug discovery. *Nat.Rev.Drug Discov.*, 1(3):211–219, 2002.

X. Pennec and N. Ayache. A geometric algorithm to find small but highly similar 3d substructures in proteins. *Bioinformatics*, 14(6):516–22, 1998.

S. J. Pickering, A. J. Bulpitt, N. Efford, N. D. Gold, and D. R. Westhead. Ai-based algorithms for protein surface comparisons. *Comput Chem*, 26(1):79–84, 2001.

A. R. Poirrette, P. J. Artymiuk, D. W. Rice, and P. Willett. Comparison of protein surfaces using a genetic algorithm. *J Comput Aided Mol Des*, 11(6):557–69, 1997.

C. P. Ponting and R. R. Russell. The natural history of protein domains. *Annu Rev Biophys Biomol Struct*, 31:45–71, 2002.

PSI. Protein structure initiative,http://www.nigms.nih.gov/psi/.

M. Rarey, B. Kramer, T. Lengauer, and G. Klebe. A fast flexible docking method using an incremental construction algorithm. *J Mol Biol*, 261(3):470–89, 1996.

M. J. Robinson, P. C. Harkins, J. Zhang, R. Baer, J. W. Haycock, M. H. Cobb, and E. J. Goldsmith. Mutation of position 52 in erk2 creates a nonproductive binding mode for adenosine 5'-triphosphate. *Biochemistry*, 35(18):5641–6, 1996.

M. Rosen, S. L. Lin, H. Wolfson, and R. Nussinov. Molecular shape comparisons in searches for active sites and functional similarity. *Protein Engineering*, 11(4):263–77, 1998.

J. Rowley. Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and giemsa staining. *Nature*, 243 (5405):290–293, 1973.

R. B. Russell. Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *J Mol Biol*, 279(5):1211–27, 1998a.

J. Saklatvala. The p38 map kinase pathway as a therapeutic target in inflammatory disease. *Curr.Opin.Pharmacol.*, 4(4):372–377, 2004.

V. Saridakis, D. Christendat, M. S. Kimber, A. Dharamsi, A. M. Edwards, and E. F. Pai. Insights into ligand binding and catalysis of a central step in nad+ synthesis: structures of methanobacterium thermoautotrophicum nmn adenylyltransferase complexes. *J Biol Chem*, 276(10):7225–32, 2001.

V. Saridakis, D. Christendat, A. Thygesen, C. H. Arrowsmith, A. M. Edwards, and E. F. Pai. Crystal structure of methanobacterium thermoautotrophicum conserved protein mth1020 reveals an ntn-hydrolase fold. *Proteins*, 48(1):141–3, 2002.

G. Scapin, S. B. Patel, J. Lisnock, J. W. Becker, and P. V. LoGrasso. The structure of jnk3 in complex with small molecule inhibitors: structural basis for potency and selectivity. *Chem Biol*, 10(8):705–12, 2003.

S. Schmitt. Dissertation. Universität Marburg, 2000.

S. Schmitt, M. Hendlich, and G. Klebe. From structure to function: A new approach to detect functional similarity among proteins independent from sequence and fold homology. *Angew. Chem. Int. Ed. Engl.*, 40(17):3141–3144, 2001.

S. Schmitt, D. Kuhn, and G. Klebe. A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.*, 323(2):387–406, 2002.

I. Schomburg, A. Chang, O. Hofmann, C. Ebeling, F. Ehrentreich, and D. Schomburg. Brenda: a resource for enzyme data and metabolic information. *Trends Biochem Sci*, 27(1):54–6, 2002a.

I. Schomburg, A. Chang, and D. Schomburg. Brenda, enzyme data and metabolic information. *Nucleic Acids Res*, 30(1):47–9, 2002b.

H. Schramek. Map kinases: from intracellular signals to physiology and disease. *News Physiol Sci*, 17:62–7, 2002.

SGC. Structural genomics consortium, http://www.sgc.ox.ac.uk/.

L. Shewchuk, A. Hassell, B. Wisely, W. Rocque, W. Holmes, J. Veal, and L. F. Kuyper. Binding mode of the 4-anilinoquinazoline class of protein kinase inhibitor: X-ray crystallographic studies of 4-anilinoquinazolines bound to cyclin-dependent kinase 2 and p38 kinase. *J Med Chem*, 43(1):133–8, 2000.

I. N. Shindyalov and P. E. Bourne. Protein structure alignment by incremental combinatorial extension (ce) of the optimal path. *Protein Eng*, 11(9):739–47, 1998.

A. Shulman-Peleg, R. Nussinov, and H. J. Wolfson. Recognition of functional sites in protein structures. *J Mol Biol*, 339(3):607–33, 2004.

T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–7, 1981.

H. Sondermann, C. Scheufler, C. Schneider, J. Hohfeld, F. U. Hartl, and I. Moarefi. Structure of a bag/hsc70 complex: convergent functional evolution of hsp70 nucleotide exchange factors. *Science*, 291(5508):1553–7, 2001.

E. L. Sonnhammer, S. R. Eddy, E. Birney, A. Bateman, and R. Durbin. Pfam: multiple sequence alignments and hmm-profiles of protein domains. *Nucleic Acids Res*, 26(1): 320–2, 1998.

M. C. Sousa and D. B. McKay. Structure of the universal stress protein of haemophilus influenzae. *Structure (Camb)*, 9(12):1135–41, 2001.

R. V. Spriggs, P. J. Artymiuk, and P. Willett. Searching for patterns of amino acids in 3d protein structures. *J Chem Inf Comput Sci*, 43(2):412–21, 2003.

M. Stahl, C. Taroni, and G. Schneider. Mapping of protein surface cavities and prediction of enzyme class by a self-organizing neural network. *Protein Eng*, 13(2):83–8., 2000.

A. Stark and R. B. Russell. Annotation in three dimensions. pints: Patterns in non-homologous tertiary structures. *Nucleic Acids Res*, 31(13):3341–4, 2003.

A. Stark, S. Sunyaev, and R. B. Russell. A model for statistical significance of local similarities in structure. *J Mol Biol*, 326(5):1307–16, 2003.

J. E. Stelmach, L. Liu, S. B. Patel, J. V. Pivnichny, G. Scapin, S. Singh, C. E. Hop, Z. Wang, J. R. Strauss, P. M. Cameron, E. A. Nichols, S. J. O'Keefe, E. A. O'Neill, D. M. Schmatz, C. D. Schwartz, C. M. Thompson, D. M. Zaller, and J. B. Doherty. Design and synthesis of potent, orally bioavailable dihydroquinazolinone inhibitors of p38 map kinase. *Bioorg Med Chem Lett*, 13(2):277–80, 2003.

C. T. Supuran. Indisulam: an anticancer sulfonamide in clinical development. *Expert Opin Investig Drugs*, 12(2):283–7, 2003.

J. Thornton. Structural genomics takes off. *Trends Biochem Sci*, 26(2):88–9, 2001.

J. M. Thornton, C. A. Orengo, A. E. Todd, and F. M. Pearl. Protein folds, functions and evolution. *J Mol Biol*, 293(2):333–42., 1999.

J. M. Thornton, A. E. Todd, D. Milburn, N. Borkakoti, and C. A. Orengo. From structure to function: approaches and limitations. *Nat Struct Biol*, 7 Suppl:991–4, 2000.

A. E. Todd, C. A. Orengo, and J. M. Thornton. Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol*, 307(4):1113–43., 2001.

P. Traxler. Tyrosine kinase inhibitors in cancer treatment (part ii). *Exp. Opin. Ther. Patents*, 8(12):1599–1625, 1998.

A. Trejo, H. Arzeno, M. Browner, S. Chanda, S. Cheng, D. D. Comer, S. A. Dalrymple, P. Dunten, J. Lafargue, B. Lovejoy, J. Freire-Moar, J. Lim, J. McIntosh, J. Miller, E. Papp, D. Reuter, R. Roberts, F. Sanpablo, J. Saunders, K. Song, A. Villasenor, S. D. Warren, M. Welch, P. Weller, P. E. Whiteley, L. Zeng, and D. M. Goldstein. Design and synthesis of 4-azaindoles as inhibitors of p38 map kinase. *J Med Chem*, 46(22):4702–13, 2003.

C. Van Kesteren, J. H. Beijnen, and J. H. Schellens. E7070: a novel synthetic sulfonamide targeting the cell cycle progression for the treatment of cancer. *Anticancer Drugs*, 13(10):989–97, 2002.

J. C. Venter et al. The sequence of the human genome. *Science*, 291(5507):1304–1351., 2001.

M. L. Verdonk, J. C. Cole, and R. Taylor. Superstar: a knowledge-based approach for identifying interaction sites in proteins. *J Mol Biol*, 289(4):1093–108, 1999.

M. L. Verdonk, J. C. Cole, P. Watson, V. Gillet, and P. Willett. Superstar: improved knowledge-based interaction fields for protein binding sites. *J Mol Biol*, 307(3):841–59, 2001.

M. Vieth, R. E. Higgs, D. H. Robertson, M. Shapiro, E. A. Gragg, and H. Hemmerle. Kinomics-structural biology and chemogenomics of kinase inhibitors and targets. *Biochim Biophys Acta*, 1697(1-2):243–57, 2004.

R. Wade and P. Goodford. Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure.

2. ligand probe groups with the ability to form more than two hydrogen bonds. *J.Med.Chem.*, 36(1):148–156, 1993.

A. C. Wallace, N. Borkakoti, and J. M. Thornton. Tess: a geometric hashing algorithm for deriving 3d coordinate templates for searching structural databases. application to enzyme active sites. *Protein Sci*, 6(11):2308–23, 1997.

A. C. Wallace, R. A. Laskowski, and J. M. Thornton. Derivation of 3d coordinate templates for searching structural databases: application to ser-his-asp catalytic triads in the serine proteinases and lipases. *Protein Sci*, 5(6):1001–13, 1996.

S. Wang, T. Sim, Y. Kim, and Y. Chang. Tools for target identification and validation. *Curr.Opin.Chem.Biol.*, 8(4):371–377, 2004.

X. Wang, X. Lin, J. A. Loy, J. Tang, and X. C. Zhang. Crystal structure of the catalytic domain of human plasmin complexed with streptokinase. *Science*, 281(5383):1662–5, 1998.

Z. Wang, P. C. Harkins, R. J. Ulevitch, J. Han, M. H. Cobb, and E. J. Goldsmith. The structure of mitogen-activated protein kinase p38 at 2.1-a resolution. *Proc Natl Acad Sci U S A*, 94(6):2327–32, 1997.

P. P. Wangikar, A. V. Tendulkar, S. Ramya, D. N. Mali, and S. Sarawagi. Functional sites in protein families uncovered via an objective and automated graph theoretic approach. *J Mol Biol*, 326(3):955–78, 2003.

J. D. Watson, A. E. Todd, J. Bray, R. A. Laskowski, A. Edwards, A. Joachimiak, C. A. Orengo, and J. M. Thornton. Target selection and determination of function in structural genomics. *IUBMB Life*, 55(4-5):249–55, 2003.

A. Weber, A. Casini, A. Heine, D. Kuhn, C. T. Supuran, A. Scozzafava, and G. Klebe. Unexpected nanomolar inhibition of carbonic anhydrase by cox-2-selective celecoxib: new pharmacological opportunities due to related binding site recognition. *J Med Chem*, 47(3):550–7, 2004.

C. H. Weber, Y. S. Park, S. Sanker, C. Kent, and M. L. Ludwig. A prototypical cytidylyltransferase: Ctp:glycerol-3-phosphate cytidylyltransferase from bacillus subtilis. *Structure Fold Des*, 7(9):1113–24, 1999.

N. Weskamp, D. Kuhn, E. Hüllermeier, and G. Klebe. Efficient similarity search in protein structure databases by k-clique hashing. *Bioinformatics*, 20(10):1522–1526, 2004.

A. Whelton. Renal aspects of treatment with conventional nonsteroidal anti-inflammatory drugs versus cyclooxygenase-2-specific inhibitors. *Am J Med*, 110 Suppl 3A:33S–42S, 2001.

A. Whelton, W. B. White, A. E. Bello, J. A. Puma, and J. G. Fort. Effects of celecoxib and rofecoxib on blood pressure and edema in patients > or =65 years of age with systemic hypertension and osteoarthritis. *Am J Cardiol*, 90(9):959–63, 2002.

C. A. Wilson, J. Kreychman, and M. Gerstein. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J Mol Biol*, 297(1):233–49, 2000.

K. P. Wilson, M. J. Fitzgibbon, P. R. Caron, J. P. Griffith, W. Chen, P. G. McCaffrey, S. P. Chambers, and M. S. Su. Crystal structure of p38 mitogen-activated protein kinase. *J Biol Chem*, 271(44):27696–700, 1996.

T. C. Wood and W. R. Pearson. Evolution of protein sequences and structures. *J Mol Biol*, 291(4):977–95, 1999.

X. Xie, Y. Gu, T. Fox, J. T. Coll, M. A. Fleming, W. Markland, P. R. Caron, K. P. Wilson, and M. S. Su. Crystal structure of jnk3: a kinase implicated in neuronal apoptosis. *Structure*, 6(8):983–91, 1998.

H. Yang, M. Yang, Y. Ding, Y. Liu, Z. Lou, Z. Zhou, L. Sun, L. Mo, S. Ye, H. Pang, G. F. Gao, K. Anand, M. Bartlam, R. Hilgenfeld, and Z. Rao. The crystal structures of severe acute respiratory syndrome virus main protease and its complex with an inhibitor. *Proc Natl Acad Sci U S A*, 100(23):13190–5, 2003.

A. Yee, K. Pardee, D. Christendat, A. Savchenko, A. M. Edwards, and C. H. Arrowsmith. Structural proteomics: toward high-throughput structural biology as a tool in functional genomics. *Acc Chem Res*, 36(3):183–9, 2003.

T. I. Zarembinski, L. W. Hung, H. J. Mueller-Dieckmann, K. K. Kim, H. Yokota, R. Kim, and S. H. Kim. Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics. *Proc Natl Acad Sci U S A*, 95(26):15189–93, 1998.

F. Zhang, A. Strand, D. Robbins, M. H. Cobb, and E. J. Goldsmith. Atomic structure of the map kinase erk2 at 2.3 a resolution. *Nature*, 367(6465):704–11, 1994.

J. Ziebuhr, E. J. Snijder, and A. E. Gorbalenya. Virus-encoded proteinases and proteolytic processing in the nidovirales. *J Gen Virol*, 81(Pt 4):853–79, 2000.

Mein herzlicher Dank gilt:

- Herrn Prof. Dr. G. Klebe für die sehr interessante Themenstellung, die Möglichkeit das rationalen Wirkstoffdesign in seiner ganzen Breite kennenzulernen sowie für die gewährte Freiheit bei der Bearbeitung des Themas.
- Herrn Prof. Dr. E. Hüllermeier für die gute Zusammenarbeit und seine Bereitschaft zur Erstellung des Zweitgutachtens.
- Dr. Stefan Schmitt für die Einführung in das Cavbase-System und in die Funktionsweise von Relibase+.
- Nils Weskamp für die sehr angenehme Form der Zusammenarbeit, seine stete Auskunftbereitschaft zu allen algorithmischen Fragestellungen sowie für zahlreiche anregende Diskussionen.
- Dr. Andreas Bergner für seine Hilfsbereitschaft und seine Auskunftbereitschaft über Relibase+.
- Dr. Alexander Weber für die angenehme Atomsphäre in A204 und die zahlreichen sehr interessanten Diskussionen.
- Paul Czodrowski für seine administrative Hilfe bei der Fertigstellung der Arbeit und für die musikalische Inspiration während der Doktorarbeit.
- Dr. Hans Matter für seine hilfreichen Anmerkungen zur Klassifizierung der Proteinkinasen.
- Meinen LiteraturarbeiterInnen Claudia Stahlschmidt und Achim Gondermann, für ihre geleistete Arbeit, deren Ergebnisse zum Teil hier eingeflossen sind.
- Allen Mitgliedern und Ehemaligen der Arbeitsgruppe Klebe für die Hilfsbereitschaft, viele interessante Diskussionen und die angenehme Arbeitsatmosphäre.

- Für das Korrekturlesen von Teilen dieser Arbeit bei Dr. Ulf Börjesson, Dr. Andreas Heine, Dr. Peter Haebel, Dr. Alexander Weber, Dr. Christoph Sotriffer, Nils Weskamp und Katharina Kuhn.
- Dr. Stefan Schmitt und Dr. Ruth Brenk für die Überlassung der LaTeX-Vorlage.
- Meinen Eltern für ihr Interesse an meiner Arbeit und für die Unterstützung während des Studiums.
- Meiner Frau Katha für ihre Inspiration und unermüdliche Unterstützung während der gesamten Dissertation, besonders während der letzten Monate.

Aus der vorliegenden Arbeit sind folgende Posterbeiträge, Vorträge und Publikationen hervorgegangen:

Tagungsbeiträge:

- Kuhn, D., Schmitt, S., Klebe, G., *Cavbase - a tool for detecting functional similarity among proteins beyond fold and sequence homology.* Poster präsentiert auf dem 16. Darmstädter Molecular Modelling Workshop, Darmstadt 2002.
- Kuhn, D. and Klebe, G., *Classification of protein families using Cavbase.* Poster präsentiert auf Fachgruppentagung Medizinische Chemie, GDCh, Fulda, September 2003.
- Kuhn, D., Weber, A., Casini, A., Heine, A., Supuran, CT., Scozzafava, A., and Klebe, G. *Cavbase: From the Successful Prediction of Cross Reactivity to the Functional Classification of Protein Families* Poster präsentiert auf dem 4. Aventis iLab Workshop, Kloster Eberbach, Januar 2004.
- Kuhn, D., Klebe, G., *Use of cavity bitstrings in the comparison of large datasets of protein cavities and the classification of protein kinases using Cavbase.* Vortrag präsentiert beim 17. Darmstädter Molecular Modelling Workshop, Erlangen 2003.
- Kuhn, D., Klebe, G., *Classification of protein kinases using Cavbase*, Vortrag präsentiert auf dem Computational Chemistry Award and Workshop, Bayer AG, Wuppertal (2003). (Der Vortrag wurde mit dem *Bayer Award for Excellence in Computational Chemistry* ausgezeichnet.)

Aufsätze:

- Schmitt, S.,Kuhn, D.,Klebe, G., *A new method to detect related function among proteins independent of sequence and fold homology*, *J. Mol. Biol.*, 2002, 18, 323(2), 387-406.

- Weber, A., Casini, A., Heine, A., Kuhn, D., Supuran, CT., Scozzafava, A., Klebe, G., *Unexpected Nanomolar Inhibition of Carbonic Anhydrase by COX-2 Selective Celecoxib: New Pharmacological Opportunities Due to Related Binding Site Recognition.*, *J. Med. Chem.* ,2004, 47, 550-7.

- Weskamp, N., Kuhn, D., Hüllermeier, E., Klebe, G., *Efficient Similarity Search in Protein Structure Databases: Improving Clique-Detection through Clique Hashing.* In H.-W. Mewes et al. (eds.): GCB 03. German Conference on Bioinformatics 2003, Proceedings - Volume I, pages 179-184. belleville Verlag, München, ISBN: 3-936-298-80-7.

- Weskamp, N., Kuhn, D., Hüllermeier, E., Klebe, G., *Efficient similarity search in protein structure databases by k-clique hashing.* Bioinformatics 20(10), pages 1522-1526, 2004.

- Weskamp, N., Hüllermeier, E., Kuhn, D., Klebe, G., *Graph Alignments: A New Concept to Detect Conserved Regions in Protein Active Sites.* In R. Giegerich, J. Stoye (eds.), Proceedings German Conference on Bioinformatics 2004, pages 131-140. GI-Lecture Notes in Informatics, Bonn, ISBN: 3-88579-382-2.

Erklärung

Ich versichere, daß ich meine Dissertation

„Beschreibung von Proteinbindetaschen für Funktionsstudien und de Novo-Design und die Entwicklung von Methoden zur funktionellen Klassifizierung von Proteinfamilien “

selbständig ohne unerlaubte Hilfe angefertigt und mich dabei keiner anderen als der von mir ausdrücklich bezeichneten Quellen bedient habe.

Die Dissertation wurde in der jetzigen oder einer ähnlichen Form noch in keiner anderen Hochschule eingereicht und hat noch keinen sonstigen Prüfungszwecken gedient.

Marburg, den 20. Dezember 2004

(Daniel Kuhn)

Lebenslauf

Daniel Kuhn

Geburtstag:	15. August 1974
Geburtsort:	Gießen
1985 – 1994	Herderschule in Gießen
1994	Allgemeine Hochschulreife
1994 – 1995	Zivildienst
Okt. 1995	Immatrikulation im Studiengang Pharmazie an der Philipps-Universität Marburg
Aug. 1997	1. Staatsexamen
Okt. 1999	2. Staatsexamen
Nov. 1999 – Apr. 2000	1. Hälfte des Praktischen Jahres in der Abteilung Chemoinformatik von der *Merck KGaA*, Darmstadt
Mai. 2000 – Okt. 2000	2. Hälfte des Praktischen Jahres in der *Burg-Apotheke* Staufenberg
Nov. 2000	3. Staatsexamen
Dez. 2000	Beginn der Arbeiten zur vorliegenden Dissertation unter Betreuung von Prof. Dr. G. Klebe am Institut für Pharmazeutische Chemie der *Philipps-Universität* Marburg
Jan. 2001 – Sep. 2004	Wissenschaftlicher Angestellter am Institut für Pharmazeutische Chemie der Philipps-Universität Marburg; Betreuung des Praktikums „Qualitative Anorganische Analyse“ (1. Semester Pharmazie)